Development of new algorithms to advance on the discovery of microRNAs

Carol Moraga

N° d'ordre NNT : 2020LYSE1222

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

École Doctorale ED 341
E2M2

Spécialité de doctorat : Bioinformatique

Soutenue publiquement le 3/11/2020, par :

Carol Moraga Quinteros

Development of new algorithms to advance on the discovery of microRNAs

---

Devant le jury composé de :

Freitas Ana Teresa, Professeur, Instituto Superior Técnico, Portugal	Rapportrice
Ossowski Stephan, Professeur, University of Tübingen, Allemagne	Rapporteur
Ruz Gonzalo A., Professeur, Universidad Adolfo Ibáñez, Chili	Rapporteur
Vieira Cristina, Professeur, Université de Lyon 1	Examinatrice
Almeida Andrea Miyasaka, Professeur, Universidad Mayor, Chili	Examinatrice
Sagot Marie-France, Directrice de recherche, INRIA	Directrice de thèse
Gutiérrez Rodrigo A., Principal Investigator, Pontificia Universidad Católica de Chile, Chili	Co-directeur de thèse
Vidal Elena A., Principal Investigator, Universidad Mayor, Chili	Co-encadrante

Por mis hijas que son todo en mi vida, y mi marido quien me inspira cada día.

UNIVERSITE CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur F. FLEURY**
Président du Conseil Académique	M. le Professeur H. BEN HADID
Vice-Président du Conseil d'Administration	M. le Professeur D. REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur P. CHEVALIER
Vice-président de la Commission Recherche	M. F. VALLÉE
Directeur Général des Services	M. A. HELLEU

COMPOSANTES SANTE

Faculté de Médecin Lyon-Est - Claude Bernard	Directeur: M. le Professeur J. ETIENNE
Faculté de Médecine et de Maeutique Lyon Sud Charles Mérieux	Directeur: Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur: M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur: Mme la Professeure C. VINCIGUERRA
Institut Techniques de Réadaptation	Directeur: M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur: Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur: M. F. De MARCHI
Département Biologie	Directeur: M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur: Mme C. FELIX
Département Génie Electrique et des Procédés	Directeur: M. Hassan HAMMOURI
Département Informatique	Directeur: M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur: M. le Professeur G. TOMANOV
Département Mécanique	Directeur: M. le Professeur H. BEN HADID
Département Physique	Directeur: M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur: M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur: M. B. GUIDERDONI
Ecole Polytechnique Universitaire de Lyon 1	Directeur: M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur: M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur: M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur: M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur: M. N. LEBOISNE

Acknowledgements

I would like to thank too many people here, because the road was rough and very beautiful at the same time, I learned a lot and not just about of my thesis's subject. Fortunately, I was not alone, I am fortunate to have many people around me constantly give me their support and love.

First to all, I thank to my family, my little girls and my husband, for being patient and stay with me all this time, they inspiring me every day of my life. Also, my family from Chile because they were always cheering me up.

I would like to thank: my advisors Marie-France Sagot, Elena A. Vidal, and Rodrigo A. Gutierrez; This thesis was possible because of your constant support, directions, and friendship, many thanks!!!. Infinite thanks to all the members of my jury who have the good and nice disposition to be part of this history, reading and attending my PhD defense, thank you very much!

Also, I want to say thanks to all the members of the Erable team, both past and present (and including those that are members of the family if not the research team); a special thanks to our team assistants Marina da Graça, Claire Sauer, and Anouchka Ronceray, for taking the burden of all the bureaucracy stuff.

Specially thanks for the fruitful conversations and coffee breaks trying to solve the world or just surviving from day to day. They were very nice moments, with a lot of laughter and many challenging Mariana's boards. I gonna miss these moments with you guys!

Thanks to the Chilean and French institutions that supported the development of this thesis: CONICYT, LBBE, INRIA, and Université Claude Bernard Lyon 1.

Thank you very much to all who were part of this exciting path in one way or another, thank you for believing in me, thank you for teaching me so much, and not just research, I learnt about life and resilience, how to be a better person every day, thanks to all of you again!

Résumé en français

Les miARNs sont de petites molécules d'ARN, plus courtes que 25 nucléotides, qui ont été identifiées comme étant des régulateurs clés de l'expression génétique au niveau post-transcriptionnel. Les miARNs sont impliqués dans un large éventail de processus biologiques, y compris le cycle cellulaire, la différenciation, l'apoptose et la pathogenèse de maladies. Il a été démontré que les miARNs pourraient également être impliqués dans la régulation inter-espèces. Chez l'homme, certains miARNs ont ainsi été identifiés comme étant des modulateurs de l'expression de gènes bactériens, suggérant qu'ils peuvent entrer dans les cellules bactériennes et réguler l'expression génétique bactérienne. Par conséquent, contrairement aux réseaux de régulation des gènes, les réseaux de régulation de miARNs peuvent être une clé pour comprendre comment les cellules hôtes peuvent interagir et réguler les millions de bactéries présentes dans l'environnement cellulaire (microbiome). De plus, les miARNs sont présents dans tous les fluides corporels et sont associés à diverses pathologies dont le cancer. Le rôle des miARNs dans la croissance et la proliferation du cancer apparaît en effet important. Les miARNs ont ainsi été identifyes comme etant des régulateurs critiques de la proliferation de nombreux types de cancers ou comme des inhibiteurs de cette prolifération. Les miARNs pourraient ainsi être des candidats potentiels pour des biomarqueurs diagnostiques, des biomarqueurs pronostiques et des cibles thérapeutiques. De plus, dans le domaine agronomique, les miARNs végétaux sont essentiels pour comprendre comment les plantes réagissent aux changements dans leur environnement et comment elles interagissent avec d'autres organismes. Par exemple, bien que les champignons pathogènes puissent causer de graves pertes aux cultures, d'autres interactions plantes-champignons améliorent la croissance des plantes, leur tolérance au stress et l'acquisition d'éléments nutritifs. Ces interactions sont essentielles à la conception de nouveaux produits biotechnologiques pour l'industrie agronomique. L'inhibition de gènes par les miARNs est un mécanisme de régulation important qui a été décrit dans les voies de défense contre les attaques de pathogènes dans divers organismes, ainsi que dans les interactions bénéfiques telles la symbiose et le mutualisme.

Il est très important de comprendre comment les miARNs communiquent et régulent l'expression au niveau du génome, mais tout d'abord, il est nécessaire de les identifier. Pour cette raison, il est essentiel de développer des algorithmes afin d'identifier les miARNs avec une grande précision, sans rien rater mais également en réduisant le nombre de faux positifs. De nos

jours, une pratique expérimentale courante consiste à capturer la séquence et les modèles d'expression des miARNs en utilisant les technologies de séquençage de nouvelle génération (NGS). De telles expériences de séquençage génèrent des millions de lectures de sARN-seq, nécessitant ainsi le développement d'algorithmes pour transformer de telles données en grande quantité en connaissances biologiques utiles. Actuellement, de nombreux outils bioinformatiques ont été développés pour analyser et identifier les miARNs mais la plupart d'entre eux s'appuient exclusivement sur les informations de conservation au niveau de la séquence et sur des génomes de référence. Les outils les plus cités sont relativement anciens, ce qui suggère une stagnation dans le développement des outils de prédiction de miARNs. Les méthodes de pointe sont fortement basées sur l'alignement des lectures de sARN-seq sur un génome de référence. De nouvelles méthodes commencent à apparaître mais restent basées sur des informations de conservation au niveau de la séquence qui ne permettent pas l'identification de nouveaux miARNs et qui sont spécifiques d'une espèce. Lorsque nous n'avons pas de génome de référence de haute qualité ou pas de génome du tout, nos possibilités sont ainsi considérablement réduites.

Cette thèse est structurée de la manière suivante. Le Chapitre 2 présente des analyses expérimentales et bioinformatiques de données doubles sRNA-seq et mRNA-seq obtenues en profilant l'interaction hôte-pathogène de *Sus scrofa* la bactérie *Mycoplasma hyopneumoniae*. L'objectif de ce travail était de démêler le réseau de régulation des miARNs orchestrant une telle interaction. Ma contribution à ce projet a été d'effectuer les analyses computationnelles pour d'abord identifier, quantifier et annoter les miARNs ainsi que d'établir un pipeline permettant la création *in silico* de réseaux de régulation miARN-ARNm à l'échelle du génome. Les résultats décrits dans ce chapitre ont été publiés dans la revue *Scientific Reports* [152], où je suis le deuxième auteur (il y deux premiers auteurs).

L'expérience que j'ai acquises dans ce travail avec les outils de pointe actuels pour la découverte de miARNs et la prédiction de leurs cibles a été essentielle pour identifier la faiblesse de ces outils et donc des lignes de recherche algorithmiques potentielles, qui se sont avérées être liées à la première étape de l'analyse des miARNs, à savoir leur identification.

Dans le Chapitre 3, je présente l'algorithme BrumiR qui est la principale contribution de cette thèse. BrumiR est un nouvel algorithme qui permet de découvrir des miARNs sans génome de référence. De plus, le chapitre présente un autre outil, miRsim qui permet de simuler des données de sRNA-seq et a été essentiel pour évaluer BrumiR sur un ensemble de données synthétiques où la vérité est connue. Les deux outils font partie de la boîte à outils BrumiR, qui est disponible gratuitement dans Github (`https://github.com/camoragaq`).

Bien que la prédiction de miARNs sans génome de référence soit utile pour les espèces non modèles, lorsqu'un génome de référence ou un projet de génome est disponible, il doit être intégré dans la découverte de miARNs.

Dans le Chapitre 4, je présente l'outil BrumiR2Reference qui peut intégrer un génome de référence pour affiner davantage les prédictions de BrumiR. Nous présentons également un benchmark de la performance de BrumiR utilisant des données publiques réelles provenant

d'espèces végétales et animales. De plus, nous démontrons l'efficacité de la boîte à outils BrumiR pour découvrir de nouveaux miARNs en utilisant des données de sRNA-seq générées à partir de racines de la plante *Arabidopsis thaliana*. Ces données ont été générées par des collaborateurs du Chili. L'outil BrumiR2reference fait également partie de la boîte à outils BrumiR et est disponible gratuitement dans GitHub (https://github.com/camoragaq/BrumiR). Les résultats présentés dans les Chapitres 3 et 4 sont décrits dans un manuscrit déjà soumis à une revue dont je suis le premier auteur. De plus, nous avons déposé notre manuscrit dans le référentiel BioRxiv (https://doi.org/10.1101/2020.08.07.240689) et tout le code de la boîte à outils BrumiR est disponible gratuitement dans GitHub (https://github.com/camoragaq).

En résumé, nous pouvons diviser les résultats de la présente thèse en deux grandes parties. La première s'est concentrée sur l'analyse de données de sARN-seq, tandis que la seconde s'est concentrée sur le développement de nouveaux algorithmes pour avancer dans la découverte de miARNs.

Contents

1 Introduction **13**

1.1 The expanding world of small RNAs. . . . 13
1.2 The era of miRNAs. . . . 14
1.3 sRNA-seq libraries. . . . 17
1.4 miRNA discovery methods. . . . 19
1.5 miRNA discovery in non-model species: a de Bruijn graph approach to organize sRNA-seq reads. . . . 22
1.6 Overview of the thesis. . . . 23

2 miRNA discovery to improve our understanding of the host-bacterium interaction between *Sus scrofa* and *Mycoplasma hyopneumoniae* **25**

2.1 Introduction. . . . 25
2.2 Implementation 27
2.2.1 Experimental procedure 27
2.2.2 Computational analysis of sequencing data. . . . 29
2.3 Results and discussion 31
2.3.1 mRNA expression profiles and differential expression 31
2.3.2 miRNA expression profiles 39
2.4 Conclusion. . . . 63

3 BrumiR algorithmn **65**

3.1 Introduction. . . . 65
3.2 Definitions. . . . 68
3.3 Implementation. . . . 68
3.3.1 Building a de Bruijn graph for sRNA-seq data. . . . 68
3.3.2 Removing sequencing errors from the unipath sRNA-seq graph. . . . 72
3.3.3 An expressed mature miRNA has uniform coverage. . . . 72
3.3.4 miRNAs and other sequences are captured in single connected components. . . . 72
3.3.5 BrumiR classifies low abundance non-linear topologies as sequencing artefacts. . . . 73
3.3.6 Re-assembling unipaths within each CC. . . . 73

3.3.7 Re-clustering potential miRNAs. 73
3.3.8 Identifying other expressed RNA sequences. 76
3.4 BrumiR algorithm: from miRNA reads to a de Bruijn graph. 78
3.5 Results: BrumiR achieves the highest accuracy on simulated data. 81
3.6 Conclusion. 86

4 Benchmarking and validating the predictions of BrumiR on a real dataset using BrumiR2reference 87
4.1 Introduction 87
4.2 Implementation 88
4.2.1 Benchmarking BrumiR using real sRNA-seq reads. 88
4.2.2 Identifying precursor sequences for the candidates of BrumiR (BrumiR2Reference) 89
4.2.3 miRNA discovery from *Arabidopsis* root samples. 90
4.3 Results 92
4.3.1 The hairpin structure of mature miRNAs is found in most of the BrumiR candidates. 92
4.3.2 Discovering novel miRNAs from sRNA-seq data of *A. thaliana* roots using BrumiR. 94
4.4 Conclusion. 105

Conclusion and Perspectives 107

Bibliography 111

Chapter 1

Introduction

1.1 The expanding world of small RNAs.

Contrary to messenger RNAs (mRNAs) which, after transcription from DNA, are translated into proteins, non-coding RNAs, also denoted by ncRNAs, are not translated into proteins but can regulate gene expression at different levels, and have essential roles in health and disease. They thus challenge the perception of molecular biology that is highly dominated by a protein-centric view [71]. The ncRNA field keeps expanding with the advent of new molecular and genomic technologies. We can see in Figure 1.1 the regulatory ncRNAs that have been identified up to now. Yet, although the accessibility of high-throughput sequencing and bioinformatic algorithms has increased our capacity to identify ncRNAs, our ability to understand and characterize the role of regulatory ncRNAs is significantly lagging behind.
Among the ncRNAs, one class is of particular interest in this thesis. This concerns what have been called small RNAs, henceforth denoted by sRNAs, that are molecules of size at most 200 nt. The first findings of sRNAs occurred in the 1990s and concerned Lin-4 and let-7 microRNAs (denoted by miRNAs) in *Caenorhabditis elegans* [159, 55].
These molecules are involved in important cellular functions at the transcriptional and/or posttranscriptional level. They include small interfering RNAs (siRNAs), piwi-interacting RNAs (piRNAs), and miRNAs that can directly regulate the expression of other RNA entities as well as their corresponding protein products; small nucleolar RNAs (snoRNAs) that guide chemical modifications (methylation, pseudouridylation); as well as other molecules such as tRNA-derived small RNAs (tsRNA), small rDNA-derived RNAs (srRNAs), and small nuclear RNAs (snRNAs). In humans, there are over 1,000 annotated miRNAs, hundreds of siRNAs, and millions of piRNAs, thus occupying a substantial portion of the genome [134]. These 3 molecules are actually present in almost all eukaryotic organisms, and have been identified in certain DNA viruses [72] and in bacteria. Among the non-coding RNAs involved in regulation, there are also long ones, such as the long non-coding RNAs denoted by lncRNAs that have been shown to be involved in the epigenetic regulation of gene expression [134].
Small RNAs, and among them the most widely studied type represented by the miRNAs will

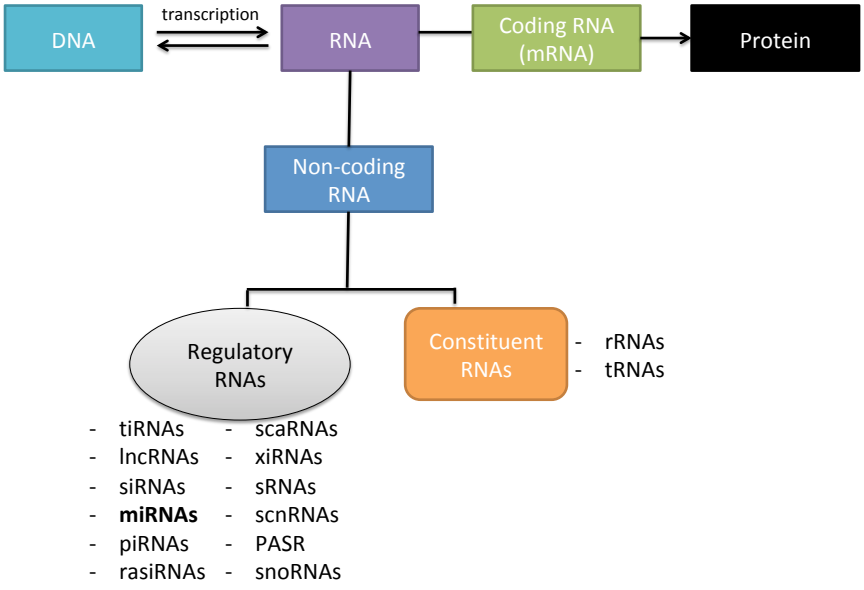

Figure 1.1 – The expanding non-coding RNA world.

be the main focus of this thesis. Notice that miRNAs are considered to be present only in eukaryotes.

1.2 The era of miRNAs.

Originally called small temporal RNAs (stRNAs), miRNAs were discovered in 1993 in *C. elegans* [108]. At the beginning, they were believed to be an oddity of nematodes. Later, it was shown that miRNAs are conserved in other metazoans, including humans [159]. In 2001, more than a hundred miRNAs had been discovered and published, some of them evolutionary conserved, thus opening a revolution in our understanding of gene regulation [100, 104, 107]. miRNAs are small RNA molecules usually shorter than 25 nucleotides (nt), which have been identified as crucial regulators of gene expression mostly at the post-transcriptional level [15]. miRNA biogenesis follows a two-step process that involves a nuclear and a cytoplasmic cleavage event [109] (Figure 1.2). In the nucleus, miRNAs are transcribed as a long primary transcript called pri-miRNA (primary miRNA). The nuclear cleavage of the pri-miRNA is done by a Class 2 ribonuclease III enzyme called Drosha, releasing a 60-150 nt stem loop intermediate called pre-miRNA (miRNA precursor sequence) that is longer in plants than in animals (average 60-70 nt). The pre-miRNA is transported into the cytoplasm where it is then cleaved by the RNase III enzyme Dicer. Such cleavage separates the loop structure, and the imperfect double strand is known as the miRNA:miRNA* duplex where miRNA

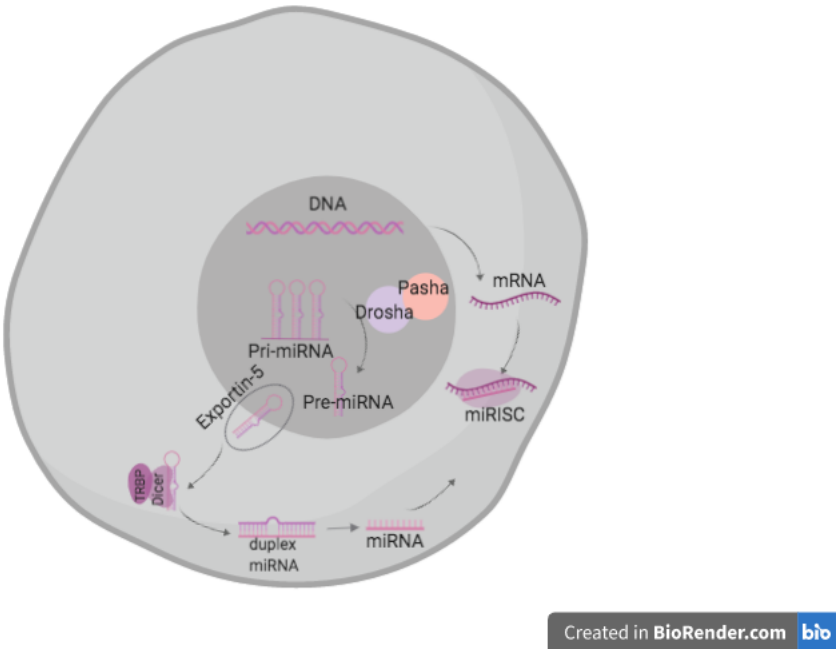

Figure 1.2 – miRNA biogenesis (Created with BioRender.com).

corresponds to a mature miRNA sequence and miRNA* or star miRNA sequence is the opposing arm of the miRNA. There are some differences in plants. For instance, the Dicer-like 1 protein has a similar function as Drosha and perhaps functions as Dicer in processing the miRNA:miRNA* duplex. Then, the duplex is transported to the nucleus by HASTY instead of by exportin 5 as is the case in animals. Both in plants and animals, the mature miRNA from the miRNA:miRNA* duplex is loaded into an RNA-induced silencing complex (RISC) and the targeting of the mRNA is done by the Argonaute protein [128].

miRNAs are involved in a wide range of biological processes including cell cycle, apoptosis and disease. During development, the expression of miRNAs occurs in a spatiotemporal, tissue and cell-specific manner, suggesting the involvement of miRNAs also in cell differentiation (see example of characterization in Zebrafish in Figure 1.3). The first evidence that miRNAs might play a role in this were in the differentiation of embryonic stem [81] and neuronal cells [106], as well as in cancer [64]. The role of miRNAs in growth and proliferation of cancer appears indeed important. miRNAs were thus identified as critical regulators of the proliferation of many cancer types [149, 215, 206] or as inhibitors of such proliferation [204]. miRNAs were also reported to play a role in cell metabolism such as lipid metabolism [203], and glucose homeostasis [163].

On the other hand, plant miRNAs play crucial roles in almost all aspects of normal plant growth and development (Figure 1.4), but also in response to environmental changes such as light, nutrition, and various abiotic and biotic stresses [27, 183, 117, 24]. Moreover, miRNAs in plants have an important role in plant-pathogen interactions. They are essential to understand how plants respond to changes in their environment and interact with other organisms. For

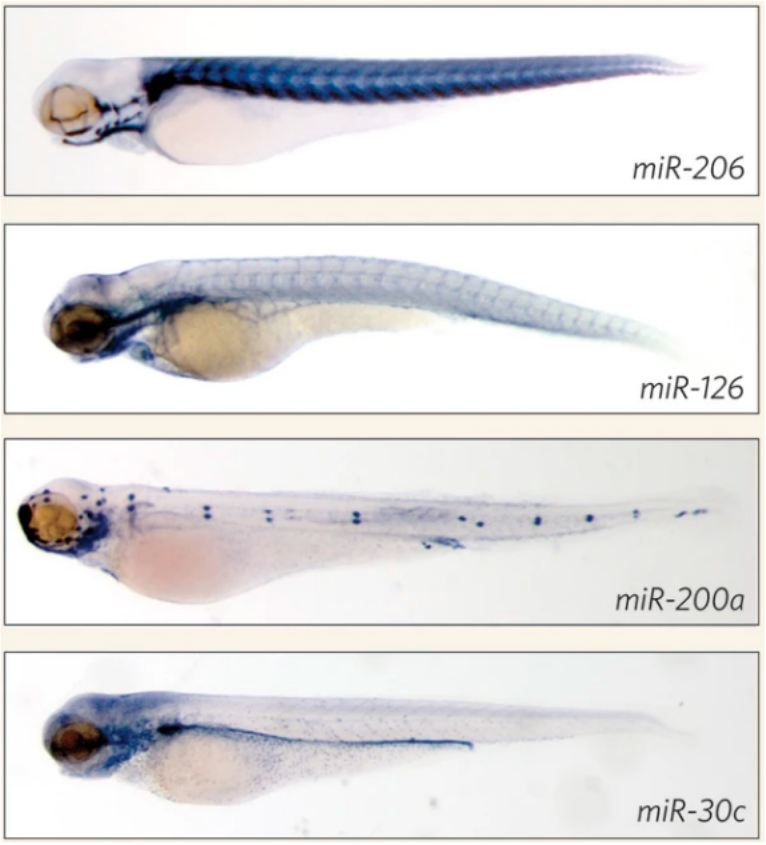

Figure 1.3 – miRNAs have tissue-specific functions in Zebrafish. Figure from [200].

Figure 1.4 – Mutant of Dicer-like 1 in *Arabidopsis thaliana* showing abnormalities in floral development produced for the silencing of miRNAs needed to regulate normal floral development [147].

instance, while pathogenic fungi may cause severe losses to crops [121], other plant-fungal interactions result in improved plant growth, tolerance to stress and nutrient acquisition [197]. These interactions are key to design new biotechnological products for the agronomic industry.

Recent evidence reveals the role of miRNAs in bacterial infectious diseases mediated by host-pathogen interactions [219]. It was thus reported that infection of gastric epithelial cells by *Helicobacter pylori* could be due to the expression changes in some miRNAs [54, 91, 216]. Although research on miRNAs in bacterial pathogen infection has greatly enhanced our understanding of these interactions, the precise mechanism underlying the regulatory function of miRNAs remains unclear [219].

More in general, even though the first discoveries of miRNAs led to an increased identification of their roles, it is the introduction of Next Generation Sequencing (NGS) technologies and their application in massive studies that opened a new way to analyze miRNAs and their characteristics [174].

1.3 sRNA-seq libraries.

Since the first classification and annotation of miRNAs, accurately identifying them as well as the regulatory networks in which they are involved has proven difficult. Accurate prediction of known and novel miRNAs along with their targets is however essential to increase our understanding of the miRNA biology. Capturing and identifying miRNAs was a challenge in the late 2010s. There were two main ways to do it: through hybridization and by sequencing

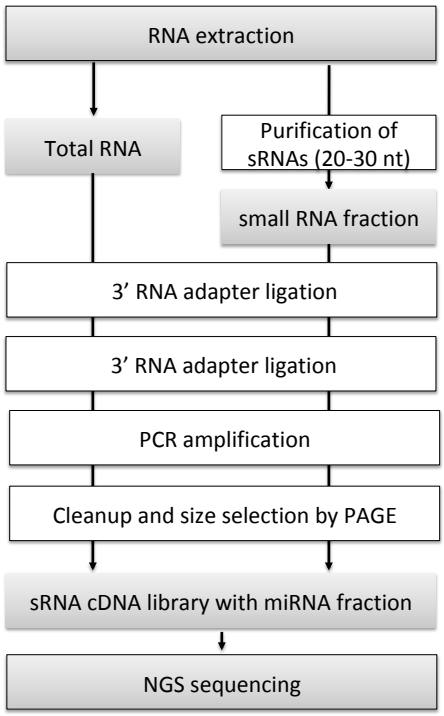

Figure 1.5 – sRNA library preparation methods for NGS platforms.

[179].

Hybridisation technologies such as microarrays allow quantification of the relative expression levels of sRNAs, but the high levels of background signal related to cross-hybridisation were a technical limitation to the detection capacity. Furthermore, the design was based on pre-existing knowledge, which prevented finding novel miRNAs [179].

Deep-sequencing technologies have opened the door to capturing and profiling known and novel miRNAs. The NGS technologies can sequence DNA orders of magnitude faster and at a lower cost in comparison to Sanger sequencing. Nowadays, a common experimental practice is to identify miRNAs and their expression patterns using such technologies [32]. Commonly, NGS experiments produce short or ultra-short reads, and a full run on an NGS platform is able to generate more than 20 million sRNA-seq reads at a relatively low cost.

Illumina Sequencing is the most current platform used to perform sRNA-seq librairies. Sample preparation is crucial for NGS experiments in order to define what biological information can be extracted and ensure the preparation of a high-quality sRNA cDNA library. The library preparation can be performed with total RNA or with enriched sRNA fraction, following on the successive ligation of 5' and 3' adapters and then a reverse transcription followed by Polymerase Chain Reaction (PCR) amplification. The final step involves sRNA cDNA gel separation and size selection by PolyAcrylamide Gel Electrophoresis (PAGE) (Figure 1.5) [127].

High-throughput RNA sequencing (sRNA-seq) has impacted the generation and discovery of new miRNA genes across the whole taxonomic tree [23]. Such information has been stored

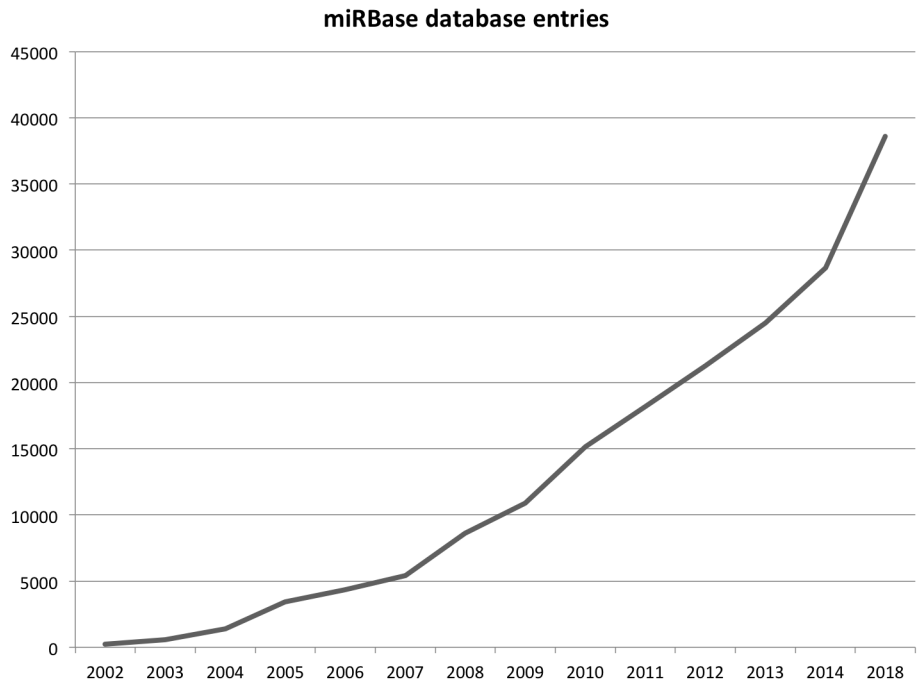

Figure 1.6 – miRBase entries along the years.

in the miRBase database [97] which is the main repository of miRNAs. The current version of miRBase stores nearly 40 thousand miRNAs. As we can see in Figure 1.6, the number of entries has sharply increased in recent years following the increasing availability of new sequencing technologies. This exponential growth along time indicates the necessity to develop new methods to process the increasing amount of available miRNA data. Many new bioinformatic tools started to appear. This is an active field of research due to the evolution of sequencing techniques [23].

1.4 miRNA discovery methods.

miRNA discovery remains one of the most important challenges in the field [32]. Nowadays, there are a large number of different tools available. Most of them are genome-based approaches, meaning that it is necessary to have a reference genome available to map the sRNA-seq reads and identify the miRNAs [32]. Furthermore, among these tools, we find different methods to make the prediction. The most popular approach is read signature evaluation which generates a set of putative precursor sequences from the reference genome and then aligns the reads against these sequences to see the distribution of the reads preserving the characteristic hairpin structure features, having a mature sequence signal and a star sequence signal [23]. Then, the excised precursor putative sequences may be evaluated using two different types of methods. Rule-base prediction methods, which are based on reference values obtained from the known precursor sequences available in miRBase [97], assign a score

to each prediction depending on the principal features that are conserved. Among such methods are MIReNA and miRDeep2 (Table 1.4). Machine-learning prediction methods on the other hand require a training step on a set of read signatures and structural features calculated from known precursor sequences as a true set, and using as a negative set random hairpin sequences. In this category of methods, we can find the tool sRNAbench (previously called miRanalyser) (see also Table 1.4) [23]. Although genome-based methods are the most popular in the field, they present the following problems: 1) when they rely exclusively on conservation information, they cannot identify species-specific miRNAs, 2) they struggle with repetitive loci because they cannot map the reads accurately in those regions which impacts the miRNA prediction, 3) they use sequence information (hairpin precursor sequence excised) from the reference genome to predict miRNAs, which is possible only when a reference genome is complete and available. In this context, *de novo* prediction methods are starting to appear as a solution when a reference genome is not available or is of low quality.

In such cases, a *de novo* approach could be complementary to the genome-based methods or be the only solution in the absence of a reference genome. To this purpose, miRNA duplex evaluation methods are a suitable option. They do not need a reference genome because they assemble the sRNA-seq reads into contigs evaluating such features as length, complementary star sequence, number of mismatches with a putative star sequence, and calculating the score using machine-learning methods that select the putative miRNA duplexes. In this group, we find the tools miReader and miRplex [23].

Our hypothesis is that it is possible to discover miRNAs by using only information contained in the sequenced reads, which was already proposed for the mirnovo tool, a machine-learning method based on a cluster analysis of the sRNA-seq reads to predict miRNAs without a reference genome based on conserved known miRNA features [195].

We believe that it is necessary to progress further in the annotation of miRNAs for non-model species by not considering also known miRNA information in order to identify new miRNA families.

Table 1.1 – Most popular miRNA discovery tools. The table displays an overview of current state-of-the-art miRNA tools. For each tool, we collected information regarding the kind of approach used, its implementation, and the short-read mapper employed. Additionally, information related to the use of external databases (miRbase), secondary structure prediction, and whether the tool can determine other types of sRNAs (isomiRs, ncRNA) is provided. Finally, information associated with the last update of the package as well as with its availability (stand-alone, web-server) is provided for each miRNA discovery tool.

tool name	organism	approach based	implementation	mapper	miRBase used	Structure prediction	isomiRs	other ncRNA	last update	platform	ref
miRPlex	A	miRNA duplex evaluation	NGS - ML	NA	yes	NA	no	no	2013	stand-alone	[135]
miReader	A, P	miRNA duplex evaluation	NGS - ML	NA	no	NA	yes	no	2016	stand-alone	[84]
sRNAbench	A, P	map to the genome - read signature	NGS - ML	Bowtie	yes	RNAfold	yes	Filter	2018	stand-alone/web server	[8]
MIReNA	A, P	map to the genome - read signature	NGS - IA	megaBlast	yes/no	RNAfold	no	filter	2013	stand-alone	[139]
miRDeep2	A	map to the genome - read signature	NGS - ML	Bowtie	yes/no	RNAfold	no	filter	2016	stand-alone	[63]
miRDeep-P	P	map to the genome - structure based	NGS - ML	Bowtie	no	NA	no	no	2011	stand-alone	[212]
miR-PREFeR	P	map to the genome - read signature	NGS - ML	Bowtie	yes/no	RNAfold			2014	stand-alone	[110]
mirnovo	A, P	de novo machine learning based on known miRNAs	NGS - ML	NA/Bowtie	yes/no	NA/RNAfold	no	filter	2017	stand-alone/web server	[195]
Mirinho	A, P	mapping to the genome - statistic approach	NGS - IA	Bowtie	yes	RNAfold	no	no	2015	stand-alone	[79]

1.5 miRNA discovery in non-model species: a de Bruijn graph approach to organize sRNA-seq reads.

Nowadays in biology in general, the study of non-model species represents a great challenge because most of the tools, databases and experimental procedures are based on model species. Furthermore, despite all the on-going work on genome assembly, we still have a low percentage of good quality reference genomes [1].

The genome assembly problem consists in building an unknown genome sequence from a set of redundant reads. The algorithms for doing this, called assemblers, attempt to reconstruct the genome by building a genome graph that encodes all the read information [155, 161]. Genome graphs are constructed from overlaps detected in the set of redundant reads. After graph construction, the genome sequence is inferred by searching the most likely genome path within the graph. The main challenge arises from repetitive elements that are often frequent within genomes and induce complex graph structures that hinder the elucidation of the genome sequence/path.

The type of genome graph constructed [155, 161] is intrinsically related to the kind of overlap detected in the set of redundant reads. Variable-length overlaps, supporting a certain degree of divergence, lead to the construction of a string graph [155], while fixed-length exact overlaps lead to the construction of a co-called *de Bruijn graph* [161]. Historically, the popularity of the latter type of graph has been dictated by the sequencing technology available to assemble a genome.

A de Bruijn graph is the preferred choice for assembling genomes from accurate short-read data because its construction avoids entirely the all-vs-all read comparison that becomes computationally intractable when we have millions of short-reads. The development of new long-read technologies rescued the string graph from oblivion, as higher error rates and longer readouts permit the computation of longer inexact overlaps that span the genome repeats and in turn reduce the complexity of the string graphs. Nowadays, short and long read technologies are commonly combined to achieve accurate and chromosome-scale genome reconstructions.

Yet, despite the current spread of long-read technologies, short-read machines still are leading by far the sequence data generation. A recent report of the European Nucleotide Archive indicates that more than 79 trillion reads and 12,546 trillion bases [6] have been stored. Interestingly, the average read length is 158 base pairs. We can infer that more than 99% of the sequence data stored in one of the largest sequence repositories have been generated using short-read technologies, which reinforces the contemporary relevance of developing algorithms based on a de Bruijn graph approach. Moreover, de Bruijn graphs have been successfully applied to assemble other kinds of omic data such as transcriptomic [68] and metagenomic [114]. Transcriptomic and metagenomic data resemble sRNA-seq data in the sense that the sequencing experiments aim to profile the abundance of a group of genes or of meta-genomes. Overall, a de Bruijn graph thus still plays a central role in genomics due to its favorable features for encoding different types of omic data.

We therefore believe that if we use a consistent structure such as a de Bruijn graph to organize the sRNA-seq reads, we can better infer the putative miRNAs.

1.6 Overview of the thesis.

This thesis is structured in the following way. In Chapter 1, I present the basic concepts to be able to follow the next chapters. In Chapter 2, I present experimental and bioinformatic analyses of dual miRNA-seq and mRNA-seq data obtained by profiling the host-pathogen interaction of *Sus scrofa* and the bacterium *Mycoplasma hyopneumoniae*. The aim of this work was to unravel the gene miRNA regulatory network orchestrating such interaction. My contribution to this project was to perform the computational analyses to first identify, quantify and annotate miRNAs as well as to build a workflow to create *in silico* miRNA-mRNA regulatory networks at genome scale. In this Chapter, I also introduce some more relevant concepts that will be key for the next chapters. The results described in this chapter were published in *Scientific Reports* [152], where I am the second author (there are two first authors).

The "on hand" experience with current state-of-the-art tools for miRNA discovery and miRNA target prediction was essential to identify the weaknesses of the current tools and therefore the potential algorithmic lines of research, which turned out to be related to the first step of miRNA analysis, namely the identification of miRNAs. In Chapter 3, I present the BrumiR algorithm, which is the main contribution of the current PhD thesis. BrumiR is a new algorithm that can discover miRNAs without a reference genome. Additionally, Chapter 3 introduced another tool, miRsim, that simulates sRNA-seq data and was key to develop and benchmark BrumiR with synthetic datasets where the ground-truth is controlled. Both tools are part of the BrumiR toolkit, which is freely available in Github (https://github.com/camoragaq).

Although, predicting miRNAs without a reference genome is useful for non-model species, when a reference or draft genome is available it should be integrated into the miRNA discovery. In Chapter 4, I present the BrumiR2Reference tool that can integrate a reference genome to further refine the miRNA predictions made by BrumiR. Additionally, we present a benchmark of the performance BrumiR using real public data from plant and animal species. Moreover, we demonstrate the effectiveness of the BrumiR toolkit for discovering novel miRNAs using sRNA-seq data generated from *Arabidopsis thaliana* roots. The latter sRNA-data and experiments were generated by collaborators from Chile. The BrumiR2reference tool is also part of the BrumiR toolkit and is freely available in GitHub(https://github.com/camoragaq/BrumiR).

The results presented in Chapters 3 and 4 are described in a manuscript already submitted to a journal, where I am the first author. Additionally, we have deposited our manuscript in the BioRxiv repository (https://doi.org/10.1101/2020.08.07.240689) and all the code of the BrumiR toolkit is freely available in GitHub (https://github.com/camoragaq).

In summary, we can divide the results of the present thesis in two big parts. The first one focused on sRNA-seq data analysis, while the second one focused on the development of new algorithms to advance on the discovery of miRNAs.

Chapter 2

miRNA discovery to improve our understanding of the host-bacterium interaction between *Sus scrofa* and *Mycoplasma hyopneumoniae*

Contents

2.1 Introduction. **25**

2.2 Implementation **27**

2.2.1 Experimental procedure 27

2.2.2 Computational analysis of sequencing data. 29

2.3 Results and discussion **31**

2.3.1 mRNA expression profiles and differential expression 31

2.3.2 miRNA expression profiles 39

2.4 Conclusion. **63**

2.1 Introduction.

Respiratory diseases are among the major health problems in the pig farming industry. *Mycoplasma hyopneumoniae* is the causative agent of swine enzootic pneumonia, a chronic respiratory disease that affects herds worldwide. Although *M. hyopneumoniae* does not cause high mortality, it is considered the most expensive pathogen for swine production [133]. This is mainly due to the costs of treatment and vaccination and to losses related to decreased

animal performance. In addition, *M. hyopneumoniae* is essential for the establishment of secondary pathogens in the host, which leads to a significant increase in mortality [132]. *M. hyopneumoniae* attaches to the cilia of the tracheal epithelial cells with participation of adhesins [48], resulting in ciliostasis and cell death [43]. Besides adhesins, virulence factors are not well understood in this bacterium. Nevertheless, a recent study from our group indicated hydrogen peroxide production from glycerol and myo-inositol metabolism as important traits that might be related with pathogenesis and with the predominance of *M. hyopneumoniae* in the swine respiratory tract [58].

MicroRNAs belong to a class of small non-coding RNAs (ncRNAs) of 18-24 nucleotides (nt) in part responsible for post-transcriptional gene regulation in eukaryotes. These evolutionarily conserved molecules influence fundamental biological processes, including cell proliferation, differentiation, apoptosis, immune response, and metabolism [15, 129]. The binding of miRNAs to target mRNAs changes the mRNA stability and translation efficiency [105], leading to degradation, suppression or up-regulation of the target mRNAs [15, 51]. Interactions between miRNA and mRNA are complex; one single miRNA can target a large number of genes belonging to diverse functional groups. Alternatively, the 3'-UTR of a single mRNA can be targeted by multiple miRNAs [120, 92]. By modulating miRNA abundance, it is thus possible to fine-tune the expression of proteins within the cell in a very precise manner [15, 92].

Recently, it was widely reported that miRNAs can be packed into exosomes and transferred to neighboring or distant cells to regulate cell function [193, 173, 31, 129]. Exosomes are small membrane vesicles (50-150 nm) released from eukaryotic cells both constitutively and upon induction, under normal and pathological conditions [173, 176]. These vesicles are involved in several cellular functions and have the potential to selectively interact with specific target cells [164, 59]. In addition to miRNAs, exosomes can transmit information among cells by transferring proteins, lipids and nucleic acids that seem to be selected non-randomly, with some specific populations of molecules being preferentially packaged into the vesicles [173, 176]. As an efficient cellular signaling and communication system, the release of exosomes by infected host cells has been recognized as a common phenomenon, in some cases beneficial to the host and in others beneficial to the pathogen [185].

Host-pathogen interactions result in signaling and physiological modifications in the host cells that induce differential miRNA expression and miRNA-mediated post-transcriptional regulation of genes involved in immune response and several other cellular pathways [154, 14]. Therefore, simultaneous identification of differentially expressed miRNAs and mRNAs provides a comprehensive view on host-pathogen interactions during the infection and the disease establishment process. In recent years, efforts have been made to identify miRNAs regulated by infection in several mammalian hosts [105, 188, 143, 14]. However, the identification of miRNAs during infection of swine cells with *M. hyopneumoniae* has not been investigated so far. To improve our understanding on the *M. hyopneumoniae*-host interaction, we sequenced and analyzed both the mRNAs and miRNAs of a swine tracheal epithelial cell line infected with *M. hyopneumoniae* strain J. In addition, we identified miRNAs differentially expressed

(DE) in the extracellular milieu and in exosome-like vesicles released by the infected cells, which play an important role in cell-cell communication and in the dissemination of host and pathogen-derived molecules during infection [176]. The simultaneous identification of miRNAs and mRNAs will help us draw a full picture of the changes in gene expression and the possible regulatory mechanisms of host cells during the disease establishment.

2.2 Implementation

2.2.1 Experimental procedure

M. hyopneumoniae strain J adhered to NPTr cells

To analyze the differential expression of New-born Pig Trachea (NPTr) cells during the infection with *M. hyopneumoniae*, we first observed the infection by immunofluorescence microscopy. These analyses were performed to detect the adherence of *M. hyopneumoniae* strain J to NPTr cells. Figure 2.1 shows the co-localization of *M. hyopneumoniae* with NPTr cells, corroborating the success of the infection. As few bacteria adhered to the cells within a short period of time (1 h vs 24 h), we chose to analyze the transcriptional alterations of NPTr cells at 24 h post-infection. *M. hyopneumoniae* strain J was chosen because previous infection assays showed that highly virulent strains, such as 7448 or 7422, damaged the host cells and we were not able to recover RNA with good sequencing quality. Although this strain is considered attenuated and incapable of causing disease *in vivo* [220, 214], our results show that *M. hyopneumoniae* is capable of adhering to the swine epithelial cells, as previously reported by Burnett *et al.* (2006) [28].

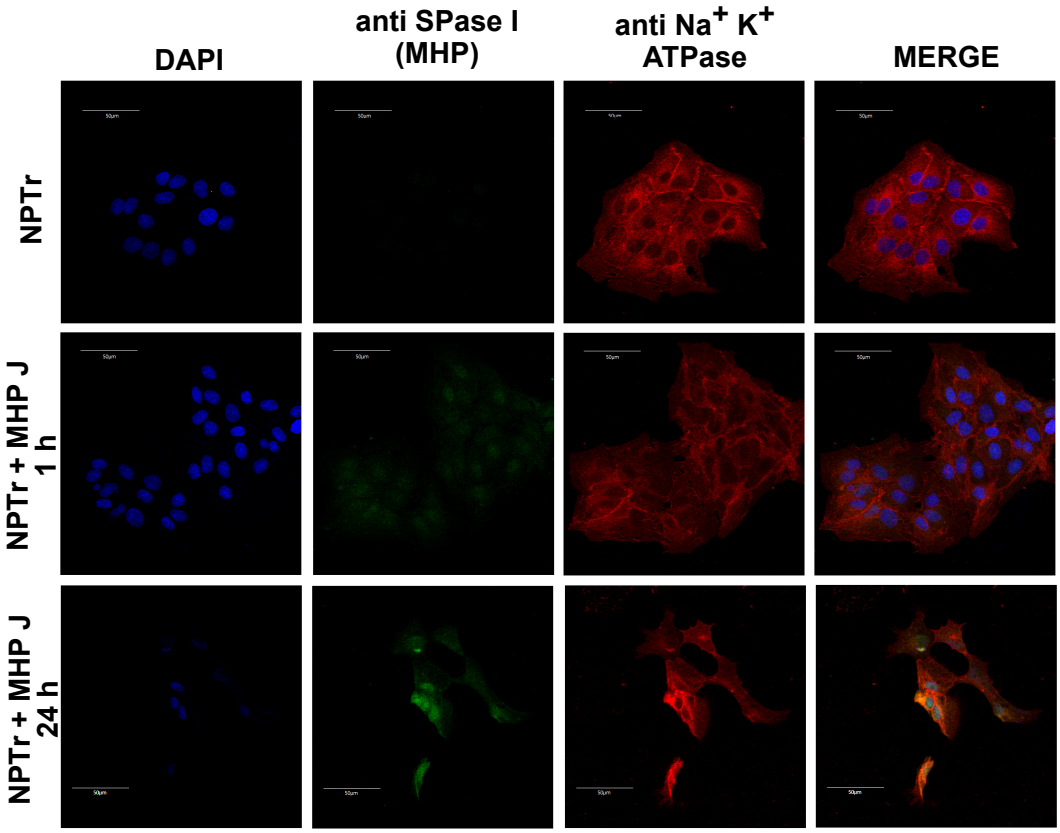

Figure 2.1 – Analysis of *M. hyopneumoniae* adherence to swine epithelial cells 1 h and 24 h post-infection. Results of the immunofluorescence microscopy indicating adherence of *M. hyopneumoniae* to the membrane of the swine cells. After 1 h few mycoplasmas were adhered to the host cells if compared to 24 h. Eukaryotic cell membranes were labeled with mouse anti-Sodium/ Potassium ATPase alpha (red), *M. hyopneumoniae* was detected with rabbit anti-SPAseI (green) and nuclei were stained with DAPI (blue). NPTr - non-infected cells. NPTr+MHP - NPTr cells infected with *M. hyopneumoniae* strain J.

Sample preparation, RNA extraction and sequencing

For mRNA sequencing, a total of 6 samples were prepared: 3 in a control group (CTL) and 3 in the infected group (INF). For miRNA sequencing, we prepared the samples as follows: total intracellular miRNA (INTRA: 3 CTL vs 3 INF), total extracellular small RNA (EXTRA: 2 CTL vs 2 INF), extracellular exosomal miRNA (EXO: 1 pool CTL vs 1 pool INF) and extracellular miRNA from vesicle-free supernatant (SN: 1 pool CTL vs 1 pool INF: a single library was constructed from a pool of 50 biological replicates). Total RNA for mRNA sequencing and sRNA enriched ($<$ 200bp) for sRNA sequencing and miRNA analyses were extracted with mirVana kit (Ambion), according to the manufacturer's instructions. Total extracellular RNA for sRNA sequencing was directly extracted from the culture NPTr cell supernatant after centrifugation of cell debris. Exosome sRNA was extracted after vesicle

purification. RNA quality was assessed with Bioanalyzer 2100 (Agilent Genomics). A total of 6 mRNA libraries were prepared (one for each sample) using TruSeq Stranded Total RNA Sample Preparation kit (Illumina) and a total of 14 small RNA libraries were prepared using TruSeq Small RNA Sample Prep kit (Illumina). After quality control, the sequencing of all libraries was performed by HiSeq2500 platform (Illumina). RNA library preparation and sequencing was performed by the Duke University GCB Sequencing Platform (Durham, USA).

2.2.2 Computational analysis of sequencing data.

Preprocessing of raw reads.

The reads from sRNA-seq and total RNA-seq were processed separately. The raw reads were filtered for low quality, and adapter sequences were trimmed with CutAdapt [137]. Total mRNA-seq clean reads were mapped against the porcine reference genome from Ensembl (Sscrofa10.2) with Star [49] and the raw counts of genes were obtained with HTSeq [7]. Only uniquely mapped reads were used for further analyses.
For miRNA prediction, clean sRNA-seq reads were mapped against the porcine genome with Bowtie [102]. We used intracellular samples as input for miRDeep2 [62] and kept predictions that had a score of at least 5. After this, we collapsed similar predictions and obtained a total of 773 clusters (773 miRNAs), of which 478 were novel miRNAs. We created a porcine miRNA DB (ssc-miRNA-DB), by including the 411 annotated porcine miRNAs in version 21 of miRBase [70, 96] along with the 722 miRNAs characterized by Martini *et al.* (2014) [138] and our novel predictions. Reads from all samples were mapped against this database and a matrix of counts was generated in order to identify DE miRNAs.

Differential expression of mRNA data

Raw counts were used as input for DE gene analysis. We detected possible DE genes in both EdgeR [171] and DESeq2 [126] packages in R. DESeq2 was run for genes that had a total count of at least 10 in all libraries, with the method's default normalization. EdgeR was used with TMM normalization and general linear model fit, only for genes with cpm greater than 1 in at least 2 libraries. After testing, the adjusted p-values (p-adj) for both methods were adjusted with the Benjamini-Hochberg [17] correction for multitesting. Genes were considered DE when p-adj < 0.05 and genes with the most pronounced logFC were selected individually for further investigation. Overall, DESeq2 detected more significant p-adj with less accentuated LogFC, while EdgeR detected less significant false discovery rates (FDR) with more extreme LogFCs. Both techniques have been widely used separately and together in several publications, and in order to select good candidates for testing, we took into account the results of both methods. For GO functional analyses, we also used the complete list of DE genes whenever LogFC was greater than 0.1 (up-regulated) or smaller than -0.1 (down-regulated). The complete lists from each method were used separately, and we compared the overall outcomes to check the robustness of our results.

Differential expression of miRNA data

The same pipeline used for DE mRNA was performed for total intracellular and total extracellular miRNAs, since we had biological replicates in both cases. In the case where we had no replicates (instead, a single library was constructed from a pool of 50 biological replicates), we used GFold [56] which provides a generalized fold change for ranking DE genes. GFold is said to overcome the shortcomings of p-value and fold change of the existing methods and can provide a more stable and biological meaningful gene ranking when a single biological replicate is available. In this case, we selected miRNAs with GFold > 2 or < -2 for functional analyses.

miRNA target prediction

DE miRNAs were used as input to detect putative interactions with the UTRs of Ensembl transcripts in the porcine genome. We used three methods to detect target pairs: miRanda [18], TargetScan [111] and Pita [90], and one method to validate the hybridization of a target pair, RNAhybrid [98]. We kept only targets that were predicted by at least two distinct tools and used the following thresholds: score in miRanda > 140, DDG from Pita < -5, score in RNAhybrid < -15. Since TargetScan does not provide a continuous scoring system, we only validated whenever there was a prediction of at least 6mers. Based on these, we chose from the target list only genes that were detected as DE in this study, and subsequentially we only considered target pairs of miRNA-mRNA that had inversed LogFC expression.

Functional analysis of DE mRNAs and targets of DE miRNAs

Functional analysis took into account as input either the list of DE mRNAs itself or the list of targets predicted for the DE miRNAs. We performed a GO enrichment analysis [11] to find out which functions were over or underrepresented in each gene list. P-values for enriched GO terms were adjusted with the Benjamini-Hochberg [17] correction for multitesting. GO terms and pathways with p-adj < 0.05 were defined as significantly enriched. The GO terms were reduced to representative non redundant terms with the use of the Revigo tool [190].

Regulatory network reconstruction and analysis

We created a general regulatory network of the host response to the bacterial infection with the DE miRNAs and target mRNAs detected in this study (Figure 2.2 and 2.1). In this network, we also included information about interactions from the BioGRID v3.4 database [189, 29], a general repository that includes experimentally validated physical and genetic interactions. We used human-based official gene symbols to include information from BioGRID, Cytoscape [178] to draw the networks, and the ClueGO plugin [19] to perform functional enrichment analysis. We further manually curated the networks for genes and miRNAs related with redox

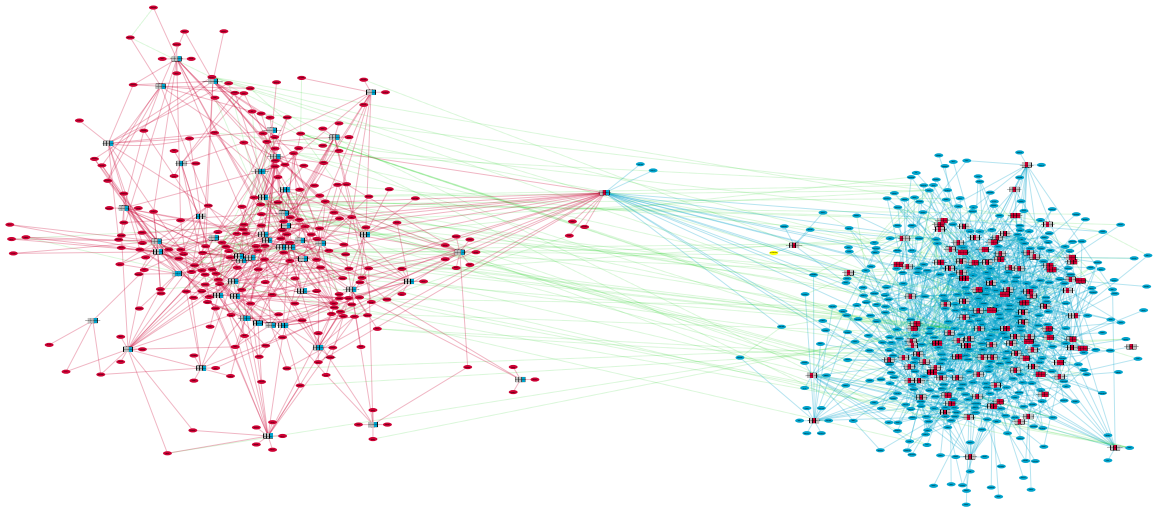

Figure 2.2 – Complete network of permissive and repressive pairs.

homeostasis (permissive regulatory network), as well as cytoskeleton and cilia (repressive regulatory network).

2.3 Results and discussion

2.3.1 mRNA expression profiles and differential expression

A total of 6 mRNA libraries were generated with the Illumina HiSeq2500 platform. A summary of the samples is provided in Table 2.1 and a diagram to explain the experimental design is given in Figure 2.3. The raw reads were submitted to the NCBI Sequence Read Archive under accession number PRJNA545822. After removing adapters and filtering low quality reads, mRNA-seq yielded around 40 million paired-end reads in all 6 samples (approx. 97% of raw reads). Trimming and mapping information for each mRNA sample is available in Table 2.2. Around 85% of the filtered reads were mapped against the porcine genome (Sscrofa10.2 - Ensembl release 89) and 40% against annotated genes (Table 2.2).

Sample ID	Sample Name	RNA Type	Compartment		Condition	Replicates
mRNA 1-3	mRNA CTRL	mRNA	Intracellular		Control	3
mRNA 4-6	mRNA INF				Infected with MHP	
sRNA 1-3	sRNA INTRA CTRL	sRNA	Intracellular		Control	
sRNA 4-6	sRNA INTRA INF				Infected with MHP	
sRNA 7-8	sRNA EXTRA CTRL		Extracellular	Total extracellular sRNA	Control	2
sRNA 9-10	sRNA EXTRA INF				Infected with MHP	
sRNA 11	sRNA EXO CTRL			sRNA from exosome-like vesicles	Control	Pool (50 biological replicates)
sRNA 12	sRNA EXO INF				Infected with MHP	
sRNA 13	sRNA SN CTRL			sRNA from vesicle-free supernatant	Control	
sRNA 14	sRNA SN INF				Infected with MHP	

Table 2.1 – **Samples detailed information.** The experimental design with the description of the samples is also explained in Figure 2.3

Sample	Total Reads (R1+R2)	Total Trimmed Read Pairs	Aligned Left	Aligned Right	Pairs Aligned	Total Alignments	Secondary Alignments	Non-Unique Alignments	Unmapped Pairs	Ambiguous Alignments	No Feature Assigned	Alignments to Genes
mRNA-1	88791632	43426720	37666684	37662008	37659633	86964046	11635354	17712828	13472366	443713	34435182	33813022
mRNA-2	88784090	43409632	37675978	37670828	37668638	86921423	11574617	17630879	13446814	440268	35251398	33041365
mRNA-3	87578734	43015169	37241617	37236499	37234429	86164756	11686640	17785255	13109876	439663	33951415	33420852
mRNA-4	87675314	42542644	36246879	36242036	36239757	84094240	11605325	17675070	15195800	478277	32366300	33005672
mRNA-5	84147016	40642408	34914105	34909268	34906777	80081455	10258082	15822818	14333462	479483	31738860	31516727
mRNA-6	85290308	41713742	36191794	36187102	36184627	83290355	10911459	16712505	12921054	480276	33114548	32452809

Table 2.2 – **Number of total, filtered and processed reads mRNA samples.**

DE Genes			DESeq2		EdgeR		UP/DOWN
Gene ID	Associated Gene Name	Description	LogFC	P-adj	LogFC	P-adj	
ENSSSCG00000028871	*LOC396866*	Cystatin-A1	0.691	4,35E-24 *	5.375	1.02E-14	Upregulated
ENSSSCG00000030385	*C3*	Complement C3	1.427	4.69E-63	2.698	2.60E-43	Upregulated
ENSSSCG00000021728	*LGALS2*	Galectin 2	0.646	5,38E-15 *	1.507	3.71E-05	Upregulated
ENSSSCG00000028525	*SAA3*	Serum Amyloid A3	0.841	4.67E-25	1.211	2.88E-21	Upregulated
ENSSSCG00000001101	*SCGN*	Secretagogin, EF-Hand Calcium Binding Protein	0.554	6.74E-09	1.018	9.06E-05	Upregulated
ENSSSCG00000028982	*AKR1C4*	Aldo-Keto Reductase Family 1 Member C4	0.869	2.61E-76	0.953	4.61E-54	Upregulated
ENSSSCG00000025277	*AKR1CL1*	Aldo-Keto Reductase Family 1, Member C-Like 1	0.866	3.96E-76	0.949	5.46E-54	Upregulated
ENSSSCG00000004678	*DUOX2*	Dual Oxidase 2	0.515	1.99E-09	0.721	6.93E-09	Upregulated
ENSSSCG00000010146	*LGALS8*	Galectin 8	0.453	1.41E-06	0.691	2.39E-15	Upregulated
ENSSSCG00000025273	*CYP11A1*	Cytochrome P450 Family 11 Subfamily A Member 1	0.51	1.39E-11	0.648	3.80E-11	Upregulated
ENSSSCG00000003402	*PGD*	Phosphogluconate Dehydrogenase	0.611	3.62E-61	0.647	2.71E-36	Upregulated
ENSSSCG00000008959	*CXCL2*	C-X-C Motif Chemokine Ligand 2	0.491	5.15E-13	0.59	1.10E-11	Upregulated
ENSSSCG00000007435	*PLTP*	Phospholipid Transfer Protein	0.436	7.49E-09	0.544	3.54E-08	Upregulated
ENSSSCG00000010853	*EPHX1*	Epoxide Hydrolase 1	0.444	5.20E-15	0.503	1.76E-06	Upregulated
ENSSSCG00000000843	*TXNRD1*	Thioredoxin Reductase 1	0.426	7.45E-14	0.482	7.33E-13	Upregulated
ENSSSCG00000017135	*ZNF750*	Zinc Finger Protein 750	-0.741	7,84E-26 *	-4.899	3.43E-15	Downregulated
ENSSSCG00000009399	*CYSLTR2*	Cysteinyl Leukotriene Receptor 2	-0.439	1,01E-07 *	-1.034	2.57E-02	Downregulated
ENSSSCG00000013401	*DKK3*	Dickkopf WNT Signaling Pathway Inhibitor 3	-0.581	5.15E-13	-0.774	1.35E-12	Downregulated
ENSSSCG00000016273	*HTR2B*	5-Hydroxytryptamine Receptor 2B	-0.434	1.37E-05	-0.735	2.12E-03	Downregulated
ENSSSCG00000014232	*LOX*	Lysyl Oxidase	-0.464	1.06E-06	-0.718	1.39E-04	Downregulated
ENSSSCG00000011217	*NEK10*	NIMA Related Kinase 10	-0.418	2.80E-05	-0.689	4.20E-03	Downregulated
ENSSSCG00000003810	*UBE2U*	Ubiquitin Conjugating Enzyme E2 U (Putative)	-0.392	1.38E-04	-0.676	8.30E-03	Downregulated
ENSSSCG00000011441	*TNNC1*	Troponin C1, Slow Skeletal And Cardiac Type	-0.399	5.92E-05	-0.633	3.27E-03	Downregulated
ENSSSCG00000015413	*FGL2*	Fibrinogen Like 2	-0.487	1.81E-15	-0.554	1.64E-14	Downregulated
ENSSSCG00000008334	*MXD1*	MAX Dimerization Protein 1	-0.385	3.21E-06	-0.486	3.02E-05	Downregulated
ENSSSCG00000027157	*SLC40A1*	Solute Carrier Family 40 Member 1	-0.365	1.23E-05	-0.462	3.19E-05	Downregulated
ENSSSCG00000006072	*VPS13B*	Vacuolar Protein Sorting 13 Homolog B	-0.359	4.30E-05	-0.469	5.52E-04	Downregulated
ENSSSCG00000028322	*BTG2*	BTG Anti-Proliferation Factor 2	-0.394	2.61E-09	-0.452	2.29E-04	Downregulated
ENSSSCG00000016059	*STAT4*	Signal Transducer And Activator Of Transcription 4	-0.354	6.36E-06	-0.43	6.75E-05	Downregulated
ENSSSCG00000028196	*MFF*	Mitochondrial Fission Factor	-0.352	6.40E-06	-0.427	3.85E-05	Downregulated

*** In the cases where EdgeR detected a DE gene with high significance and the padj was not calculated in DESeq2 due to the presence of one outlier, we checked case by case and validated as DE, whenever suitable. The adjusted P-value showed in these cases is the value calculated prior to multitest for demonstration purposes only.**

Table 2.3 – **Selected up- and down-regulated genes.** Information about the top 15 up- and 15 down-regulated DE genes from both the DeSeq2 and EdgeR methods. Ordered by LogFC calculated by EdgeR.

We detected a total of 20,274 (out of 23,215) genes expressed with at least 10 counts across the 6 mRNA libraries from NPTr cells. Then a DE analysis was performed using DESeq2 and EdgeR. We detected 1,268 DE genes (p-adj < 0.05, 517 up-regulated and 751 down-regulated), from which 502 were common to two well known methods for the detections of DE genes, 721 were exclusive to DESeq2 and 45 were exclusive to EdgeR. Information from the top 15 up-regulated and down-regulated DE genes is provided in Table 2.3. The results of gene ontology (GO) enrichment analysis for up-regulated and down-regulated genes in DESeq2 are shown separately in Figure 2.4 (p-adj < 0.05 and absolute fold enrichment ≥ 0.1). For the up-regulated genes, the enriched terms either in biological process (BP), molecular function (MF) or cellular component (CC) were related to protein synthesis (translation/ribosome), oxidation reduction activity and cell-cell communication (such as exosomes, anchoring and focal adhesion functions) (Figure 2.4A). For the down-regulated genes, the majority of the overrepresented terms in all three GO categories were related to cell cycle, cell division, cilia and cytoskeleton (Figure 2.4B).

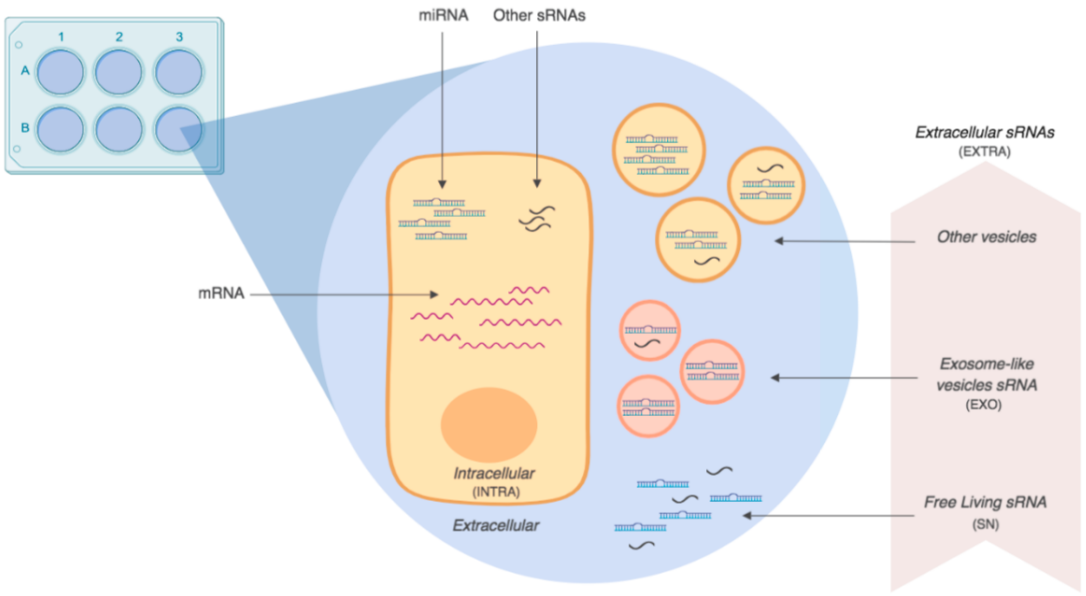

Figure 2.3 – Experimental design of samples. Eukaryotic cells express mRNAs and miRNAs and also export some of these molecules to the extracellular environment in vesicles or in free form. Therefore, besides analyzing differential expression of intracellular mRNA samples (mRNA), we also analyzed DE miRNAs in distinct cell compartments. In this way, we extracted sRNAs from: i) cells (INTRA); ii) from the medium of cultured cells, which contained all kinds of vesicles released by the cells (microvesicles, exosomes, apoptotic bodies, etc) as well as free-living sRNAs (EXTRA); iii) from exosome-like vesicles released by cells (EXO); and iv) from the supernatant of the ultracentrifugation of exosome-like vesicles, which contained free-living miRNAs (SN). The sRNAs of all these sources were sequenced and the differential expression between infected and non infected samples was analyzed.

M. hyopneumoniae elicited an antioxidant response and induced the accumulation of NRF2 in the nuclei of NPTr cells

Among the up-regulated genes, we found several ones related to immune response and inflammation, such as C3 complement, *SAA3*, chemokines (*CXCL2* and *CCL20*) and galectins (*LGALS2* and *LGALS8*). Interestingly, we also detected 64 up-regulated genes related to redox homeostasis and antioxidant defense (Table 2.4). We observed that 46 out of 64 have already been described as targets of the nuclear factor erythroid 2-related factor 2 (NRF2) in closely related species (Table 2.4). This transcription factor is involved in the protection of the cell against oxidative damage through transcription activation of cytoprotective genes [37, 130]. More specifically, several studies have shown the protective role of NRF2 in bacterial lung infections in rodents, being a critical factor for assembling the innate immune response in the host [192, 12, 167, 67].

Indeed, *M. hyopneumoniae* infection induced the expression of genes related to the two biggest redox systems in NPTr cells - glutathione (*GGT1, GGT5, GGLC, GSR*) and thioredoxin (*TXNRD1, PRDX5, PRDX6*) -, and also genes coding for NADPH-regenerating enzymes (used by the aforementioned redox systems), as analogously reported for activation of NRF2 targets in mice [89]. Moreover, a number of antioxidant genes and genes coding for detoxification enzymes (such as the *AKR* gene family, *NQO1, HMOX* and *GST*) which were up-regulated during *M. hyopneumoniae* infection, were also reported to be activated by NRF2 [89] (Table 2.4). We performed a Fisher's exact test to compare the proportion of genes within the genome expected to be targets of NRF2 with the proportion of up-regulated genes putatively regulated by this transcription factor. The results indicate an extremely significant correlation of an NRF2 target being an upregulated gene (p-value < 0.00001, Table 2.5). The up-regulation of several targets of NRF2 was further validated by reverse-transcription quantitative PCR (RT-qPCR) and the results are available in Figure 2.5.

One of the many compounds known to activate the NRF2 pathway is hydrogen peroxide [60]. The production of hydrogen peroxide was previously described to have an important cytotoxic effect in several *Mycoplasma* species, such as *M. pneumoniae* [75] and *M. mycoides* subsp. *mycoides* [194]. More specifically, the cytotoxity of *M. mycoides* subsp. *mycoides* was correlated to the bacterium's ability to translocate hydrogen peroxide directly into the host cell [20]. Previously, we have identified that *M. hyopneumoniae* was capable of producing this toxic product from glycerol metabolism [58]. Although the production of hydrogen peroxide by the *M. hyopneumoniae* strain J was not detected in that work, novel analyses indicate that in conditions where glucose is scarce, the attenuated strain J is able to produce this toxic metabolite (Figure 2.6). In addition, we were able to detect hydrogen peroxide in the medium of the NPTr cells infected with *M. hyopneumoniae* in the presence of glycerol (Figure 2.6.C). Although the production was higher in the cells infected with the pathogenic strain 7448, *M. hyopneumoniae* strain J was also able to produce the toxic product, indicating

that both strains could potentially cause oxidative damage to the host cells. These results are in accordance with the gene expression of the putative enzyme responsible for hydrogen peroxide production (GlpO) in *M. hyopneumoniae* strain J, which did not differ from the pathogenic strains [58]. During infection in general, bacteria need to compete with host cells (and other organisms in the environment) for glucose and other energy sources. For instance, in *M. pneumoniae*, Halbedel *et al.* (2004) [74] showed that even though glucose was the most efficient carbon source for biomass yield (as is the case for *M. hyopneumoniae*), the authors propose that glycerol is not only a carbon source, but it could be used by this species as an indicator that they reached their preferred ecological niche, a lipid-rich mucosal surface. Thus, it is plausible to say that *M. hyopneumoniae* also uses glycerol as a carbon source *in vivo*, however whether this is a result of competition against other fast-growing bacteria or if this species is targeting a glycerol-rich niche, we cannot affirm at this point. What we can affirm is that glycerol is not the most efficient carbon source for both *M. pneumoniae* and *M. hyopneumoniae*, and yet these species make use of it for energy uptake and also for other advantages, such as hydrogen peroxide production. Thus, it is possible that the hydrogen peroxide produced by *M. hyopneumoniae* strain J as a result of glycerol metabolism might be one of the triggers that activated the transcription factor NRF2.

It has been demonstrated in other species that NRF2 is largely regulated at the post-transcriptional level by its sub-cellular distribution, which is controlled by the Kelch-like ECH-associated protein (KEAP1). Under normal conditions, a portion of NRF2 is retained in the cytoplasm via its interaction to KEAP1 and it is subsequently ubiquitinated and degraded by the proteasome. In response to oxidative stress, reactive cysteines in KEAP1 are modified, generating conformational changes in the complex and releasing NRF2, which is translocated and accumulates into the nucleus [89, 21, 130]. In our analysis, we did not detect a difference in the expression of the mRNAs of NRF2 (base mean expression around 8000 reads, a logFC between infected and non-infected conditions close to zero and a non-significant adjusted p-value of 0.99) and KEAP1 (base mean expression around 4500 reads, logFC between infected and non-infected conditions close to zero and non-significant adjusted p-value of 0.94). However, we were able to demonstrate by confocal immunofluorescence microscopy (Figure 2.7) that both the attenuated and the virulent strains of *M. hyopneumoniae* induced a statistically significant accumulation of this transcription factor in the nuclei of NPTr cells (Figure 2.7B). In the nucleus, NRF2 is able to recognize and bind to antioxidant response element (ARE) motifs in the promoter region of target genes, activating their transcription [157]. In this work, we detected at least one conserved ARE sequence upstream the start codon in the promoter regions of 44 out of the 46 NRF2 predicted targets (with stringent search to TG/TAnnnnGC) with the use of fuzznuc software from EMBOSSv6.6.0 package [170].

In this way, it seems that *M. hyopneumoniae* infection has the potential to cause oxidative stress to the host cells, which in turn activate antioxidant response genes induced by NRF2 to fight the infection and maintain cellular homeostasis. We believe that this oxidative stress was in part related to the hydrogen peroxide produced by this bacterium, although more

experiments are needed to prove the association between this mechanism and the up-regulation of antioxidant response genes.

M. hyopneumoniae induced down-regulation of cytoskeleton and ciliary genes as well as a decrease of actin stress fibers in NPTr cells

We identified several down-regulated genes related to ciliary function, cytoskeleton and cell cycle/cell division. The impairment of the ciliary motility is a well-known effect caused by several *Mycoplasma* respiratory species [9], such as *M. pneumoniae* and *M. gallisepticum* [2, 30]. It is also well established that *M. hyopneumoniae* attaches to cilia of epithelial cells and promotes ciliostasis and loss of cilia, causing damage to the mucociliary apparatus [43, 214]. Therefore, our hypothesis is that one of the reasons for epithelial damage, besides physical adhesion, could be associated with modulations in gene expression induced by the infection, which is a running hypothesis for at least two epithelial pathogens from the genus *Mycobacterium* [142]. In agreement with this hypothesis, we were able to identify the down-regulation of genes coding for axonemal dyneins (*DNAH11, DNAH12, DNAI2, DNAL1*), which are essential for the ciliary motility [94]. In addition, genes necessary for axonemal dynein assembly (*DYX1C1*) [191], genes related to ciliogenesis (*CEP162, DCDC2, MACF1, IFT57*)[198, 177, 144, 180], ciliary polarization (*INTU*) [213], ciliary beating (*MYO1D*) [78] and several others in which the mutation or knockout is associated with ciliopathies (*LRRC6, MNS1, AK7*) [57, 218, 83] were down-regulated in the infected cells (Table 2.6). Interestingly, in line with our results, a recent study compared the transcriptional response of unvaccinated and vaccinated chicken infected with *M. gallisepticum* and the authors identified enrichment of GO terms in down-regulated genes related to cilia and cytoskeleton in unvaccinated animals [10]. Protein functions encoded by the top down-regulated genes were involved in microtubule assembly and stability, axonemal dynein complex assembly, and formation and motor movement of cilia, indicating that at least in one *Mycoplasma* species the ciliary damage caused by infection could be also explained by the down-regulation of genes involved in the ciliary function.

Besides ciliary genes, we also detected the down-regulation of cytoskeleton-related genes, both from microtubules and actin filaments, involved in the organization, rearrangement and stability of these structures (Table 2.6). It was previously described that the intracellular species *M. penetrans* is able to trigger reorganization of the host cell cytoskeleton, promoting aggregation of tubulin and α-actinin and condensation of phosphorylated proteins [66]. To investigate if *M. hyopneumoniae* indeed affected the host cell cytoskeleton, we verified the organization of actin fibers in infected cells by confocal immunofluorescence microscopy. The actin stress fibers were notably less evident in the infected cells, as opposed to the control condition, in which they were abundant and more evenly distributed (Figure 2.8). However, whenever present in the infected conditions, these stress fibers were either disorganized and/or

at the periphery of the cells. These results corroborate studies from Raymond *et al.* (2018) [165], which suggested that this species induces cytoskeletal rearrangements in the porcine respiratory tract. Although actin is not a major component of the ciliary axoneme, actin cytoskeleton has been implicated in every stage of ciliogenesis and many aspects of ciliary function [148], directly associating these two down-regulated functions. In addition, it has recently been shown that *M. hyopneumoniae* expresses surface-accessible actin-binding proteins and that the host's extracellular actin may act as a receptor for this bacterium in PK-15 epithelial cells [166], indicating its importance for successful infection. Furthermore, the reduction of visible actin stress fibers caused by *M. hyopneumoniae* may also be related to the activation of NRF2, since the actin cytoskeleton is a scaffold necessary to maintain the transcription factor in the cytoplasm [88].

We also identified the down-regulation of cytoskeleton-related genes that play a role during cell division, such as *BUB1B, CENP-I, NEK2* and *SPAG5*. Many of these genes are involved with the mitotic spindle and chromosome segregation [131, 93, 202, 22, 201, 207, 140, 196]. Genes encoding microtubule dependent motor proteins that physically affect chromosome segregation, such as kinesins (usually up-regulated during mitosis) [38], as well as genes related to cell cycle progression were down-regulated during infection. These results suggest a repression of cell division of infected cells. The manipulation of cell cycle by pathogens has been extensively reported, with different pathogens being able to arrest different points of the cell cycle [182, 136, 85, 3]. Therefore, it is possible that *M. hyopneumoniae* infection also interferes with the host cell cycle.

2.3.2 miRNA expression profiles

A total of 14 small RNA (sRNA) libraries were generated with the Illumina HiSeq2500 platform. A complete description of the samples is provided in Table 2.1 and Figure 2.3. The raw reads were submitted to the NCBI Sequence Read Archive under accession number PRJNA545822. After removing adapters and filtering low quality reads, sRNA-seq yielded from 1 to 21 million single-end clean reads. Trimming and mapping information for each sRNA sample is given in Table 2.7.

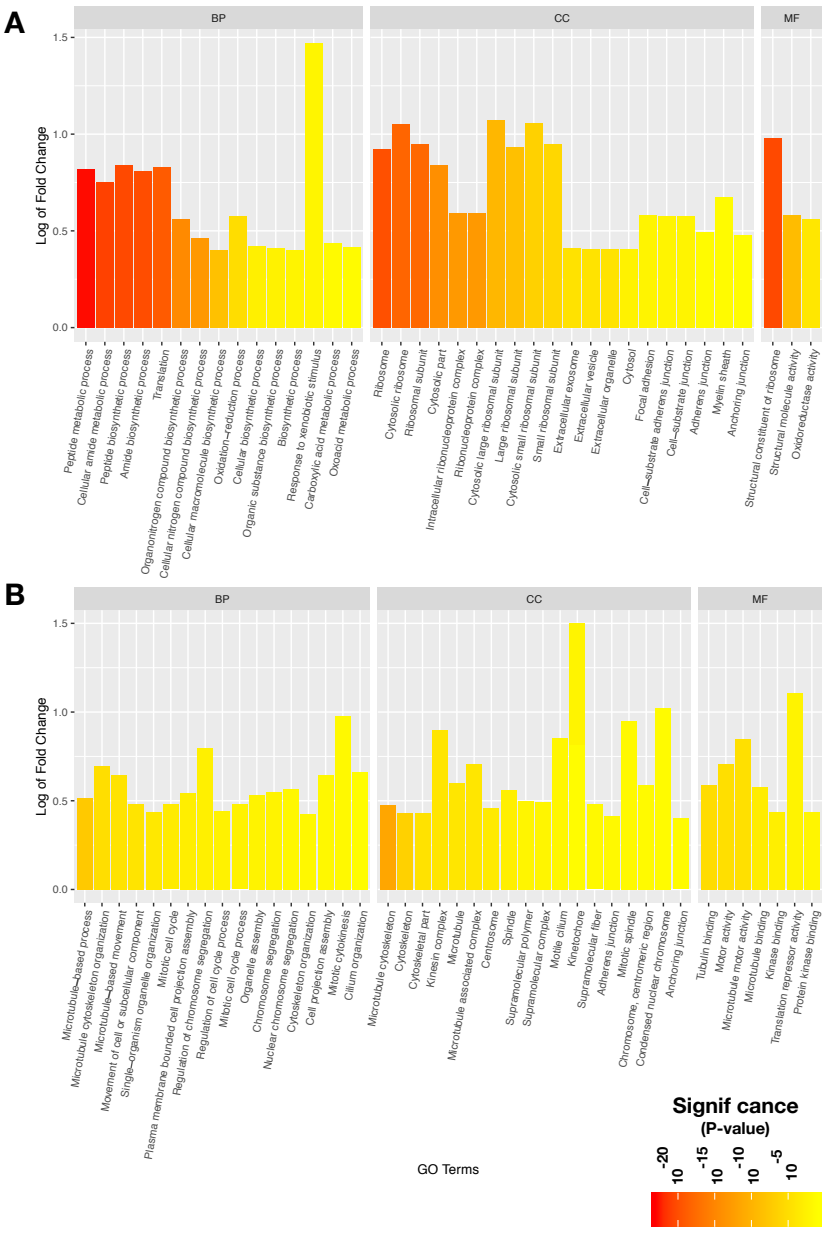

Figure 2.4 – **Significantly enriched Gene Ontology functions of DE mRNAs. A.** Up-regulated genes were enriched in terms related to protein synthesis, oxidation-reduction activity and cell-cell communication (such as exosomes, anchoring and focal adhesion), either in biological process (BP), molecular function (MF) or cellular component (CC). **B.** Down-regulated genes showed an enrichment of terms related to cytoskeleton, ciliary function, cell cycle and cell division in all the three categories of GO.

DE Genes			DESeq2		EdgeR		Reference[1]
Gene ID	Associated Gene Name	Description	LogFC	padj	LogFC	FDR	
ENSSSCG00000028982	*AKR1C4*	Aldo-Keto Reductase Family 1 Member C4	0.869	2.61E-76	0.953	4.61E-54	Hayes & McMahon, 2009 [77]
ENSSSCG00000025277	*AKR1CL1*	Aldo-Keto Reductase Family 1, Member C-Like 1	0.866	3.96E-76	0.949	5.46E-54	Hayes & McMahon, 2009[77]
ENSSSCG00000003402	*PGD*	Phosphogluconate Dehydrogenase	0.611	3.62E-61	0.647	2,71E-36	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000025273	*CYP11A1*$^{+}$	Cytochrome P450 Family 11 Subfamily A Member 1	0.510	1.39E-11	0.648	3.80E-11	
ENSSSCG00000010853	*EPHX1*	Epoxide Hydrolase 1	0.444	5.20E-15	0.503	1.76E-06	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000014540	*FTH1*	Ferritin Heavy Chain 1	0.436	1.81E-24	0.470	1.85E-14	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000003153	*FTL*	Ferritin Light Chain	0.427	1.25E-24	0.459	5.31E-14	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000000843	*TXNRD1*	Thioredoxin Reductase 1	0.426	7.45E-14	0.482	7.33E-13	Hayes & Dinkova-Kostova, 2014[76]
ENSSSCG00000002754	*NQO1*	NAD(P)H Quinone Dehydrogenase 1	0.415	7.89E-18	0.456	1.64E-12	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000002626	*GSTA1*$^{+}$	Glutathione S-Transferase A1	0.407	1.02E-12	0.461	4.18E-09	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000016312	*UGT1A6*$^{+}$	UDP Glucuronosyltransferase Family 1 Member A6	0.367	1.88E-18	0.394	7.49E-12	Hayes & Dinkova-Kostova, 2014[76]
ENSSSCG00000011147	*AKR1C1*	Aldo-Keto Reductase Family 1 Member C1	0.360	1.28E-04	0.511	1.20E-03	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000021386	*PTGR1*	Prostaglandin Reductase 1	0.358	6.02E-12	0.397	1.55E-04	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000001488	*GCLC*	Glutamate-Cysteine Ligase Catalytic Subunit	0.350	2.78E-10	0.392	2.89E-09	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000002825	*CES1*$^{+}$	Carboxylesterase 1	0.348	5.58E-12	0.384	2.93E-08	Hayes & Dinkova-Kostova, 2014[76]
ENSSSCG00000012136	*PIR*	Pirin	0.334	9.34E-07	0.392	7.21E-06	Brzska *et al.*, 2014 [26]
ENSSSCG00000025762	*GSR*$^{+}$	Glutathione S-Reductase	0.314	2.46E-03	0.489	2.57E-02	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000029144	*SRXN1*$^{+}$	Sulfiredoxin 1	0.309	3.31E-06	0.359	1.12E-05	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000010055	*GGT5*	Gamma-Glutamyltransferase 5	0.305	1.38E-03	0.421	3.33E-03	
ENSSSCG00000014833	*UCP2*	Uncopling Protein 2	0.303	5.10E-03	0.509	4.52E-02	
ENSSSCG00000028099	*SLC6A6*$^{+}$	Solute Carrier Family 6 Member 6	0.295	2.36E-03	0.412	5.50E-03	Hayes & McMahon, 2009 [77]
ENSSSCG00000026759	*HMOX1*	Heme Oxygenase 1	0.272	9.35E-03	0.402	4.36E-02	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000008311	*CYP26B1*	Cytochrome P450 Family 26 Subfamily B Member 1	0.267	4.82E-03	0.359	8.22E-03	
ENSSSCG00000028996	*ALDH1A1*	Aldehyde Dehydrogenase 1 Family Member A1	0.261	1.47E-18	0.274	2.92E-07	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000006717	*PHGDH*	Phosphoglycerate Dehydrogenase	0.260	1.20E-07	0.285	4.62E-05	
ENSSSCG00000021067	*BLVRB*	Biliverdin Reductase B	0.259	2.00E-04	0.302	1.22E-03	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000013366	*LDHA*	Lactate Dehydrogenase A	0.255	9.07E-07	0.281	8.19E-05	
ENSSSCG00000029781	*SELENOM*	Selenoprotein M	0.254	2.25E0-2	0.406	5.78E-02	
ENSSSCG00000001723	*PLA2G7*	Phospholipase A2 Group VII	0.250	4.58E-04	0.294	2.13E-03	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000004093	*IYD*	Iodotyrosine Deiodinase	0.250	1.09E-02	0.339	1.83E-02	
ENSSSCG00000001963	*EGLN3*	Egl-9 Family Hypoxia Inducible Factor 3	0.248	7.36E-03	0.324	9.69E-03	
ENSSSCG00000003914	*PRDX1*	Peroxiredoxin 1	0.231	1.02E-07	0.249	1.27E-04	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000017092	*GPX3*	Glutathione Peroxidase 3	0.224	7.69E-05	0.250	1.89E-03	Kensler *et al.*, 2007 [89]
ENSSSCG00000018048	*SLC5A10*	Solute Carrier Family 5 Member 10	0.215	3.21E-03	0.251	9.13E-03	Hayes & McMahon, 2009 [77]
ENSSSCG00000003491	*AKR7A2*	Aldo-Keto Reductase Family 7 Member A2	0.205	1.69E-02	0.252	3.10E-02	Li *et al.*, 2015 [113]
ENSSSCG00000021408	*TKT*	Transketolase	0.190	2.39E-06	0.204	1.49E-03	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000010340	*FAM213A*	Family With Sequence Similarity 213 Member A	0.184	1.62E-05	0.198	4.20E-03	
ENSSSCG00000010056	*GGT1*	Gamma-Glutamyltransferase 1	0.178	1.95E-02	0.209	2.50E-02	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000013030	*PRDX5*	Peroxiredoxin 5	0.177	1.04E-02	0.204	2.22E-02	Graham *et al.*, 2018 [69]
ENSSSCG00000012847	*TALDO1*	Transaldolase 1	0.177	3.29E-03	0.199	2.33E-01	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000030461	*HEPHL1*	Hephaestatin Like 1	0.170	8.19E-05*	3.003	3.60E-04	
ENSSSCG00000018084	*ND3*	Mitochondrially Encoded NADH:Ubiquinone Oxidoreductase Core Subunit 3	0.166	2.19E-03	0.183	2.77E-02	
ENSSSCG00000012327	*HSD17B10*	Hydroxysteroid 17-Beta Dehydrogenase 10	0.164	1.66E-02	0.188	1.94E-02	
ENSSSCG00000003007	*BCKDHA*	Branched Chain Keto Acid Dehydrogenase E1, Alpha Polypeptide	0.163	3.52E-02	0.191	5.54E-02	
ENSSSCG00000010928	*KDM5B*	Lysine Demethylase 5B	0.161	2.36E-04	0.174	1.58E-02	
ENSSSCG00000014336	*EGR1*	Early Growth Response 1	0.159	2.52E-04	0.172	1.42E-02	Gomez *et al.*, 2016 [67]
ENSSSCG00000008550	*SLC5A6*	Solute Carrier Family 5 Member 6	0.155	2.36E-03	0.170	2.44E-02	Hayes & McMahon, 2009.
ENSSSCG00000006324	*ALDH9A1*	Aldehyde Dehydrogenase 9 Family Member A1	0.155	6.71E-04	0.168	2.07E-02	
ENSSSCG00000022742	*PRDX6*	Peroxiredoxin 6	0.151	1.49E-03	0.164	2.28E-02	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000001701	*HSP90AB1*	Heat Shock Protein 90 Alpha Family Class B Member 1	0.149	2.11E-04	0.160	2.15E-02	Hayes & McMahon, 2009 [77]
ENSSSCG00000028739	*CEBPB*	CCAAT/Enhancer Binding Protein Beta	0.148	1.32E-02	0.165	3.70E-01	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000004454	*ME1*	Malic Enzyme 1	0.145	5.30E-03	0.160	2.96E-01	Hayes & Dinkova-Kostova, 2014 [76]
ENSSSCG00000010682	*PRDX3*	Peroxiredoxin 3	0.139	1.35E-02	0.154	7.06E-02	Gomez *et al.*, 2016 [67]
ENSSSCG00000027525	*DHCR24*	24-Dehydrocholesterol Reductase	0.123	3.68E-03	0.133	5.52E-02	
ENSSSCG00000029876	*SOD2*	Superoxide Dismutase 2	0.112	1.66E-02	0.123	1.39E-01	Reszka *et al.*, 2013 [169]

[1]References related to genes regulated by NRF2.
$^{+}$Genes with more than one ID identified as DE.
* The P-value showed in these cases is the value calculated prior to multitesting for demonstration purposes only (see Table 2).

Table 2.4 – **Up-regulated genes involved in redox homeostasis.** Detailed information for differential expression of genes (from both the DeSeq2 and EdgeR methods) related to antioxidant and redox homeostasis functions. The great majority of such genes has been demonstrated to be activated by the NRF2 transcription factor in related species. Duplicated entries related to different transcripts of a given gene were deleted in this table, however calculation of the statistics was made taking into account the total amount of 64 genes (including duplications).

	Upregulated	Not Upregulated	Marginal Row Totals
Targets of NRF2	46	604	650
Not Targets	172	22393	22565
Marginal Column Totals	218	22997	23215

Table 2.5 – NRF2 contingency Table. * The Fisher exact test statistic value is less than 0.00001. * Chi-square with Yates correction Chi squared equals 264.082 with 1 degrees of freedom. The two-tailed P value is less than 0.0001 The association between rows (groups) and columns (outcomes) is considered to be extremely statistically significant. * Chi-square without Yates correction Chi squared equals 270.827 with 1 degrees of freedom. The two-tailed P value is less than 0.0001 The association between rows (groups) and columns (outcomes) is considered to be extremely statistically significant.

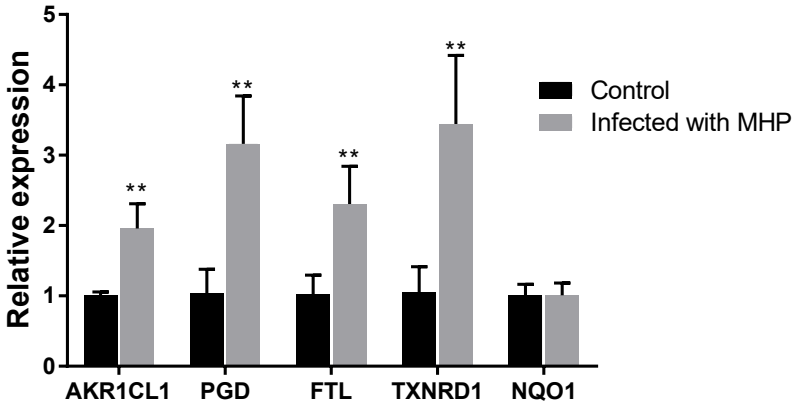

Figure 2.5 – RT-qPCR of selected NRF2 targets. Up-regulated genes described to be activated by the transcription factor NRF2 were selected for experimental validation by RT-qPCR. Of these, four of them (AKR1CL1, PGD, FTL and TXNRD1) were in accordance with the expression of sequencing data. NQO1 showed no different expression in RT-qPCR , in contrast with the up-regulation observed in the sequencing data. (** $p<0.01$).

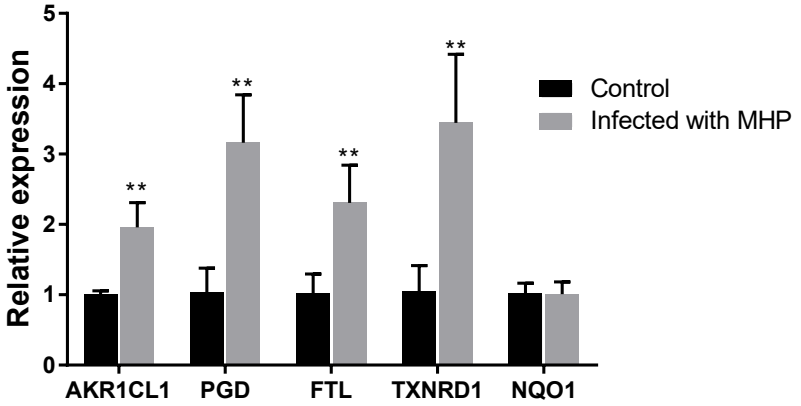

Figure 2.6 – Production of hydrogen peroxide by different *M. hyopneumoniae* strains. A) In defined medium after bacterial growth: Hydrogen peroxide was slightly detected in growth media from the attenuated strain J. Data are presented as mean and standard deviation of three independent samples. B) In the presence of different carbon sources: Pathogenic and attenuated strains of *M.hyopneumoniae* were used to test hydrogen peroxide production in incubation buffer supplemented with either glycerol or glucose after 10 2 h of incubation. All strains were able to produce significant amounts of the toxic product when glycerol was present. C) In the medium of NPTr cells after 24 h of infection with *M. hyopneumoniae* in the presence of glycerol. Data are represented as mean and standard deviation of four independent samples (* $p<0.05$; **** $p<0.0001$).

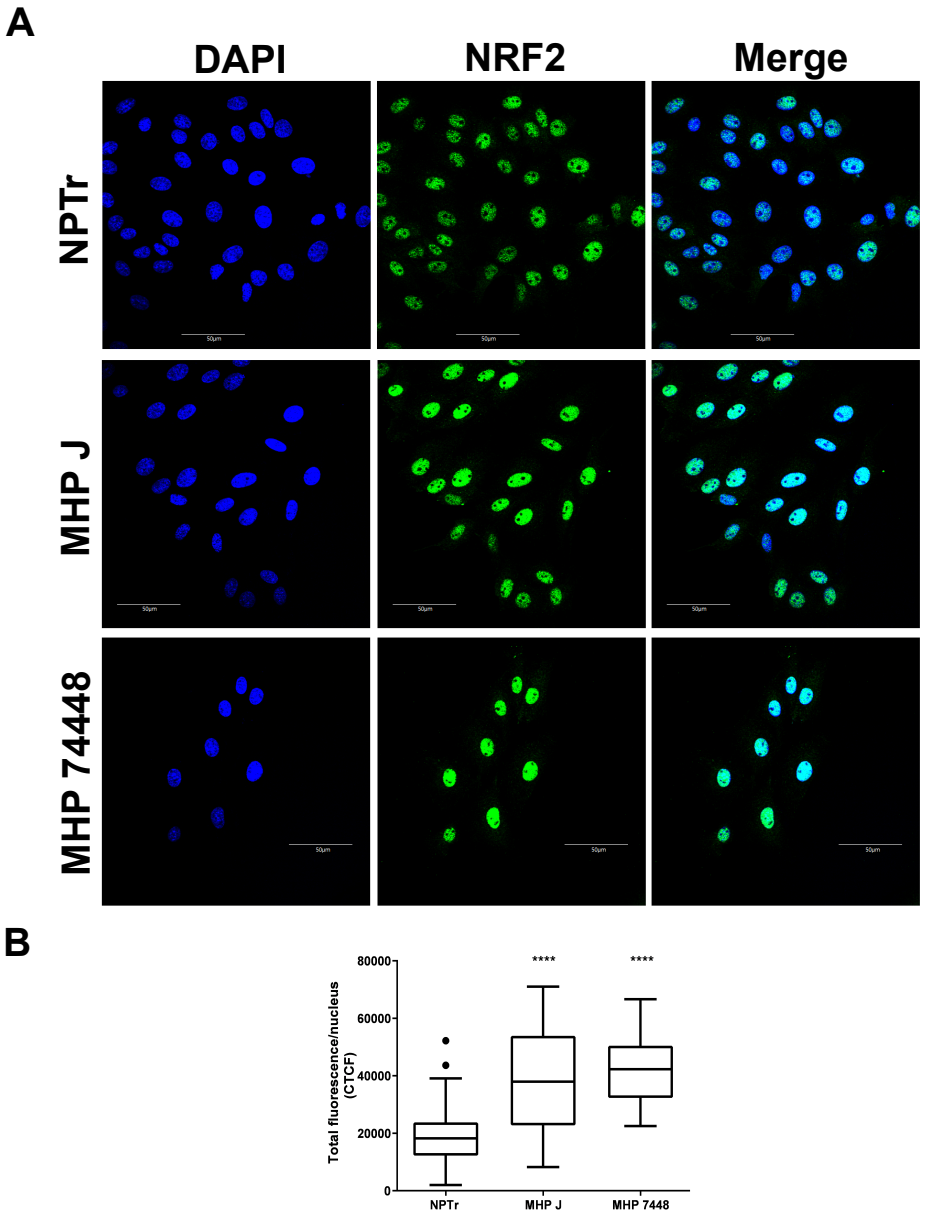

Figure 2.7 – **Localization and expression pattern of the NRF2 protein in cells infected with *M. hyopneumoniae*.** **A.** Results of the immunofluorescence microscopy analysis indicating the accumulation of NRF2 in the nuclei of infected cells after 1 h of incubation. Both attenuated (J) and virulent (7448) strains of *M. hyopneumoniae* induced the accumulation of the transcription factor in the nuclei of the epithelial cells. Eukaryotic cell NRF2 was labeled with anti-NRF2 antibody (green) and nuclei were stained with DAPI (blue). Scale bars: 50 μm. **B.** Total NRF2 green fluorescence per nucleus of control and infected cells. Both *M. hyopneumoniae* strains significantly increased the concentration of NRF2 in the nuclei of infected cells. Boxplots represent the measures of at least 25 nuclei in each condition. Outliers are represented as black dots. CTCF - corrected total cell fluorescence. NPTr - uninfected cells. MHP J - NPTr cells infected with *M. hyopneumoniae* strain J. MHP 7448 - NPTr cells infected with *M. hyopneumoniae* strain 7448. **** $p < 0.0001$

DE Genes			DESeq2		EdgeR	
Gene ID	**Associated Gene Name**	**Description**	**LogFC**	**P-adj**	**LogFC**	**P-adj**
ENSSSCG00000011441	***TNNC1***	Troponin C1, Slow Skeletal And Cardiac Type	-0.399	5.92E-05	-0.633	3.27E-03
ENSSSCG00000001791	***SAXO2***	Stabilizer Of Axonemal Microtubules 2	-0.328	2.08E-03	-0.547	4.38E-02
ENSSSCG00000014191	***FER***	FER Tyrosine Kinase	-0.331	1.75E-04	-0.426	4.07E-04
ENSSSCG00000015036	***DIXDC1***	DIX Domain Containing 1	-0.352	3.48E-07	-0.405	2.57E-06
ENSSSCG00000009079	***INTU***	Inturned Planar Cell Polarity Protein	-0.273	9.94E-03	-0.403	4.79E-02
ENSSSCG00000015531	***CEP350***	Centrosomal Protein 350	-0.314	3.28E-04	-0.400	2.58E-03
ENSSSCG00000024357	***DNAI2***	Dynein Axonemal Intermediate Chain 2	-0.276	7.42E-03	-0.395	4.04E-02
ENSSSCG00000002347	***DNAL1***	Dynein Axonemal Light Chain 1	-0.285	2.87E-03	-0.378	8.99E-03
ENSSSCG00000004705	***MAP1A***	Microtubule Associated Protein 1A	-0.271	7.62E-03	-0.376	2.25E-02
ENSSSCG00000001728	***CD2AP***	CD2 Associated Protein	-0.330	5.69E-08	-0.366	8.97E-05
ENSSSCG00000000874	***GAS2L3***	Growth Arrest Specific 2 Like 3	-0.270	5.65E-03	-0.361	1.70E-02
ENSSSCG00000010896	***ASPM***	Abnormal Spindle Microtubule Assembly	-0.291	3.90E-05	-0.332	3.79E-02
ENSSSCG00000015329	***PPP1R9A***	Protein Phosphatase 1 Regulatory Subunit 9A	-0.245	1.54E-02	-0.328	3.17E-02
ENSSSCG00000016542	***LRGUK***	Leucine Rich Repeats And Guanylate Kinase Domain Containing	-0.230	2.63E-02	-0.309	5.49E-02
ENSSSCG00000016725	***TNS3***	Tensin 3	-0.239	1.28E-02	-0.306	2.94E-02
ENSSSCG00000016608	***IQUB***	IQ Motif And Ubiquitin Domain Containing	-0.248	2.68E-03	-0.295	1.01E-02
ENSSSCG00000002780	***TPPP3***	Tubulin Polymerization Promoting Protein Family Member 3	-0.251	1.52E-03	-0.294	8.99E-03
ENSSSCG00000006394	***CFAP45***	Cilia And Flagella Associated Protein 45	-0.251	1.56E-03	-0.294	6.49E-03
ENSSSCG00000013332	***KIF18A***	Kinesin Family Member 18A	-0.227	2.00E-02	-0.290	3.88E-02
ENSSSCG00000021571	***KIF27***	Kinesin Family Member 27	-0.246	1.78E-03	-0.287	9.15E-03
ENSSSCG00000020990	***DNAH12***	Dynein Axonemal Heavy Chain 12	-0.235	6.82E-03	-0.283	1.75E-02
ENSSSCG00000000530	***FGD4***	FYVE, RhoGEF and PH domain containing 4	-0.231	7.32E-03	-0.277	3.30E-02
ENSSSCG00000003931	***KIF2C***	Kinesin Family Member 2C	-0.253	4.30E-05	-0.276	3.27E-02
ENSSSCG00000015523	***RALGPS2***	Ral GEF With PH Domain And SH3 Binding Motif 2	-0.255	1.06E-06	-0.270	3.30E-02
ENSSSCG00000008768	***ARAP2***	ArfGAP With RhoGAP Domain, Ankyrin Repeat And PH Domain2	-0.220	7.54E-03	-0.256	3.11E-02
ENSSSCG00000006936	***ODF2L***	Outer Dense Fiber Of Sperm Tails 2 Like	-0.213	1.57E-02	-0.255	4.25E-02
ENSSSCG00000001869	***PEAK1***	Pseudopodium Enriched Atypical Kinase 1	-0.210	1.58E-02	-0.250	3.89E-02
ENSSSCG00000011941	***IFT57***	Intraflagellar Transport 57	-0.223	3.97E-04	-0.242	2.12E-03
ENSSSCG00000011745	***PRKCI***	Protein Kinase C Iota	-0.225	5.70E-05	-0.239	8.27E-02
ENSSSCG00000002504	***AK7***	Adenylate Kinase 7	-0.209	2.77E-03	-0.231	4.20E-03
ENSSSCG00000017884	***TEKT1***	Tektin 1	-0.199	1.40E-02	-0.229	5.40E-02
ENSSSCG00000005952	***LRRC6***	Leucine Rich Repeat Containing 6	-0.199	1.28E-02	-0.227	3.64E-02
ENSSSCG00000005078	***DAAM1***	Dishevelled Associated Activator Of Morphogenesis 1	-0.206	1.97E-03	-0.225	8.79E-03
ENSSSCG00000009629	***BIN3***	Bridging Integrator 3	-0.195	1.39E-02	-0.221	3.88E-02
ENSSSCG00000015379	***DNAH11***	Dynein Axonemal Heavy Chain 11	-0.204	1.68E-03	-0.221	3.21E-03
ENSSSCG00000016658	***ANLN***	Anillin Actin Binding Protein	-0.212	1.05E-05	-0.220	1.89E-03
ENSSSCG00000016794	***MYO10***	Myosin X	-0.210	6.68E-05	-0.220	3.93E-03
ENSSSCG00000004969	***KIF23***	Kinesin Family Member 23	-0.205	6.08E-05	-0.214	4.68E-02
ENSSSCG00000003654	***MACF1***	Microtubule-Actin Crosslinking Factor 1	-0.201	4.70E-04	-0.214	1.08E-02
ENSSSCG00000015570	***IVNS1ABP***	Influenza Virus NS1A Binding Protein	-0.205	1.05E-05	-0.212	1.91E-02
ENSSSCG00000004289	***CEP162***	Centrosomal Protein 162	-0.191	1.01E-02	-0.212	3.17E-02
ENSSSCG00000010457	***KIF20B***	Kinesin Family Member 20B	-0.192	1.92E-03	-0.206	4.54E-02
ENSSSCG00000010351	***CCSER2***	Coiled-Coil Serine Rich Protein 2	-0.201	2.27E-06	-0.206	2.70E-03
ENSSSCG00000007235	***TPX2***	TPX2, Microtubule Nucleation Factor	-0.189	3.76E-04	-0.198	4.85E-02
ENSSSCG00000017728	***MYO1D***	Myosin ID	-0.184	5.05E-05	-0.189	2.22E-02
ENSSSCG00000010471	***KIF11***	Kinesin Family Member 11	-0.179	5.69E-04	-0.186	2.09E-02
ENSSSCG00000016781	***TRIO***	Trio Rho Guanine Nucleotide Exchange Factor	-0.174	1.51E-03	-0.182	2.87E-02
ENSSSCG00000009348	***STARD13***	StAR Related Lipid Transfer Domain Containing 13	-0.174	4.69E-04	-0.180	2.09E-02
ENSSSCG00000009793	***CLIP1***	CAP-Gly Domain Containing Linker Protein 1	-0.174	1.19E-04	-0.178	9.64E-03
ENSSSCG00000001085	***DCDC2***	Doublecortin Domain Containing 2	-0.162	2.19E-03	-0.168	3.95E-02
ENSSSCG00000005235	***KANK1***	KN Motif And Ankyrin Repeat Domains 1	-0.163	2.54E-04	-0.166	5.13E-02
ENSSSCG00000009125	***ANK2***	Ankyrin 2	-0.159	4.57E-03	-0.166	4.54E-02
ENSSSCG00000010313	***VCL***	Vinculin	-0.159	1.43E-05	-0.159	1.20E-02
ENSSSCG00000016061	***MYO1B***	Myosin IB	-0.150	1.13E-03	-0.152	5.52E-02

Table 2.6 – **Down-regulated genes involved in cytoskeleton and cilia.** Detailed information for differential expression of genes (from both the DeSeq2 and EdgeR methods) related to cytoskeleton and ciliary functions.

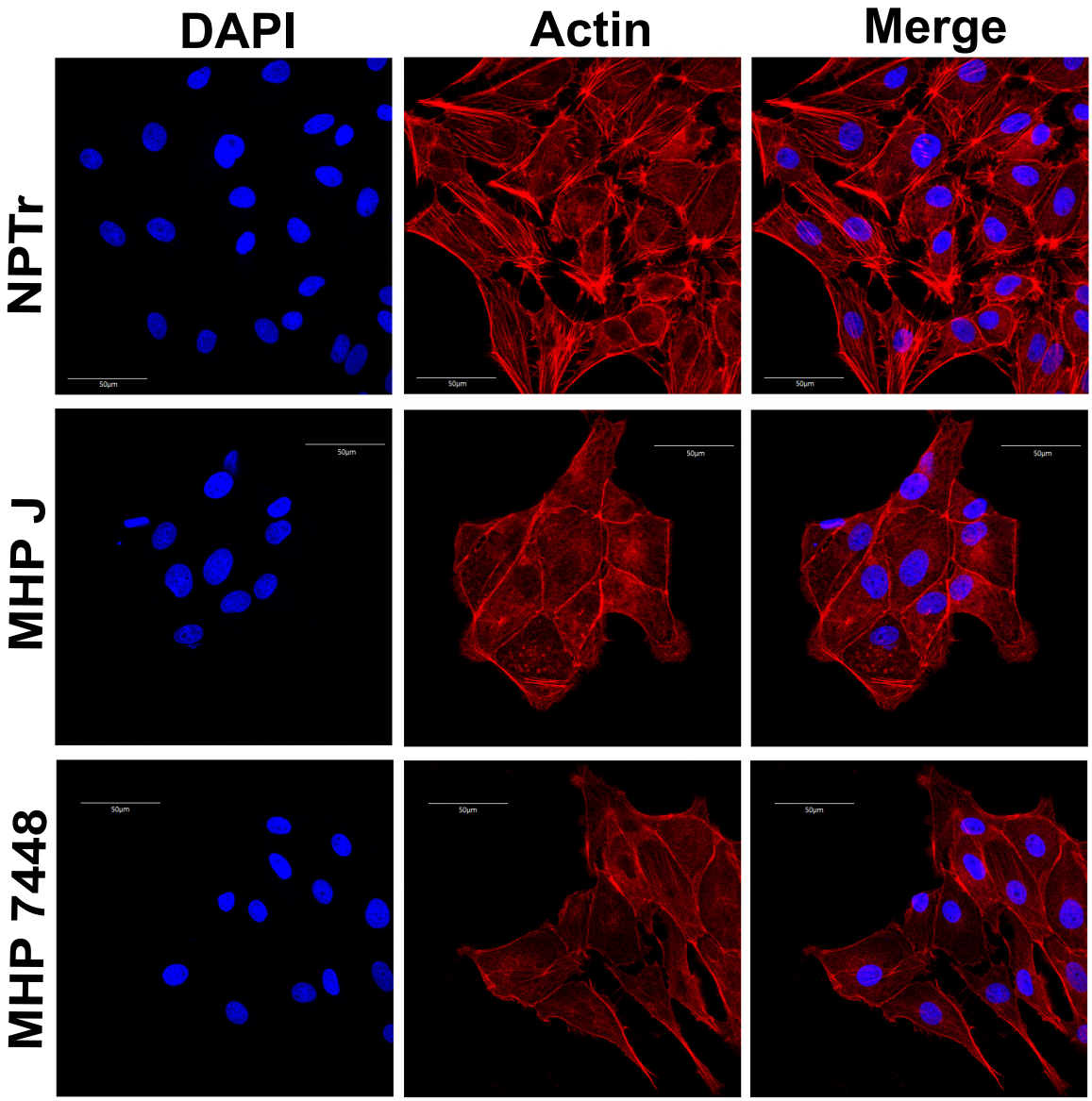

Figure 2.8 – **Organization of actin fibers in cells infected with *M. hyopneumoniae*.** Results of the immunofluorescence microscopy analysis indicating the reduction and change in the pattern of actin stress fibers in infected cells. Eukaryotic cell actin was labeled with phalloidin (red) and nuclei were stained with DAPI (blue). Both attenuated (J) and virulent (7448) strains of *M. hyopneumoniae* altered the organization and abundance of actin fibers after 24 h of infection, however this effect can already be seen after 1 h of incubation (not shown). NPTr - uninfected cells. MHP J - NPTr cells infected with *M. hyopneumoniae* strain J. MHP 7448 - NPTr cells infected with *M. hyopneumoniae* strain 7448. Scale bars: 50 μm.

Samples	Raw Reads	trimmed Reads	Reads Between 18-25nt (miRNA)	Map to S.scrofa genome (kept)	Map to S.scrofa tscp	Map to M. hyopneumoniae genome	Map to ssc-miRNA-DB	Map to M. hyopneumoniae
	Count	Count	Count	Count	Count	Count	Count	Count
sRNA-1	14798320	13562448	11870834	12759929	11739853	416	7055741	NA
sRNA-2	14057425	12749066	11001803	11868453	10890353	1029	6504845	NA
sRNA-3	14364858	13508797	11154312	12680657	11383819	2618	6584122	NA
sRNA-4	15174413	13752658	11043222	12650527	11013724	6228	6343176	NA
sRNA-5	16469852	14847785	13439426	13990427	13049626	191	8103208	NA
sRNA-6	19539863	15635652	13132448	14642718	12967407	2134	7793082	NA
sRNA-7	8444491	1092912	427861	1024395	627722	533	44189	NA
sRNA-8	918430	381103	92005	298214	143246	36	9580	NA
sRNA-9	21071682	10450248	4119935	8128629	4642956	2069111	253017	2115
sRNA-10	12469883	5335069	2197135	4018715	2534491	1236287	73012	1727
sRNA-11	12980440	6935695	2424142	6672960	3975305	719	419381	NA
sRNA-12	6839321	2682450	1071552	2477114	1553261	13232	437282	NA
sRNA-13	14404242	10321210	4080677	10051163	1610673	641	20490	NA
sRNA-14	1174960	37086	6558	24223	19256	136	1110	NA
Average	12336299	8663727	6147279	7949152	6153692	238094	3117303	NA

Table 2.7 – **Number of raw, filtered, processed, and mapped reads from miRNA samples.**

For intracellular sRNAs (INTRA: samples S1 to S6), we were able to map around 93% of the reads against the porcine genome. We also mapped all sRNA reads against the *M. hyopneumoniae* strain 7448 genome and samples had no more than 0.05% of unique mycoplasmal reads (Table 2.7). All intracellular sRNA samples had similar sequencing depth and showed a typical miRNA distribution length curve ranging from 18 to 26 nt, with a peak at 22 nt (Figure 2.9.A). We also performed a homology analysis of raw reads and we provide the percentage of intracellular clean reads clustered by classical types of sRNAs (Figure 2.9.B), showing an enrichment of miRNA-homolog reads (around 40%).

As expected, total extracellular sRNA samples (EXTRA: S7 to S10) had more RNA degradation due to the presence of RNAses and degraded mRNAs in the extracellular environment. Extracellular sRNA samples of cells infected with the bacterium (S9 and S10) had up to 20% of reads mapped to *M. hyopneumoniae*, as they were presumed to have more remnants of mycoplasmal cells (Table 2.7). However, the great majority of reads mapping to the bacteria were mainly product of mRNA degradation, as they did not have any specific signature for sRNAs when compared to previous results from Siqueira *et al.* (2016) [186]. In this way, we filtered out (i) sequences $<$ 18 nt and (ii) sequences mapped to the *M. hyopneumoniae* genome. Sample S8 (EXTRA) had a sequencing depth smaller than its replicate (S7); however, the distribution of counts was overall similar between replicates and we included it in the further analyses.

Extracellular exosomal sRNA samples (EXO: S11 and S12) had some RNA degradation ($<$18 nt), but maintained a pronounced peak at 22 nt. Extracellular sRNAs from vesicle-free supernatant (SN: samples S13 and S14) were more problematic, probably due to (i) too many pre-processing steps and (ii) the presence of RNAses in the extracellular environment. Sample S14 did not yield a minimum amount of reads necessary for the subsequent steps and this condition (SN) was not used for further analyses.

Annotation of known and novel miRNAs

In order to perform miRNA prediction with miRDeep2 [62], we only took into account reads from intracellular samples. We predicted a total of 1,041 miRNAs, which were further clustered into 773 groups. From these, 478 were completely novel

We created a *Sus scrofa* miRNA database (ssc-miRNA-DB) with 1,906 different entries from three different sources: 411 known miRNAs in *Sus scrofa* from miRBase (release 21) [96], 722 annotated by Martini *et al.*, (2014) [138] and the 773 clusters of mature miRNAs predicted by our analysis. More than 50% of intracellular sRNA clean reads were aligned against the ssc-miRNA-DB (Table 2.7). All other samples were also mapped against this database. The complete pipeline used in this study for miRNA prediction is described in Figure 2.10.

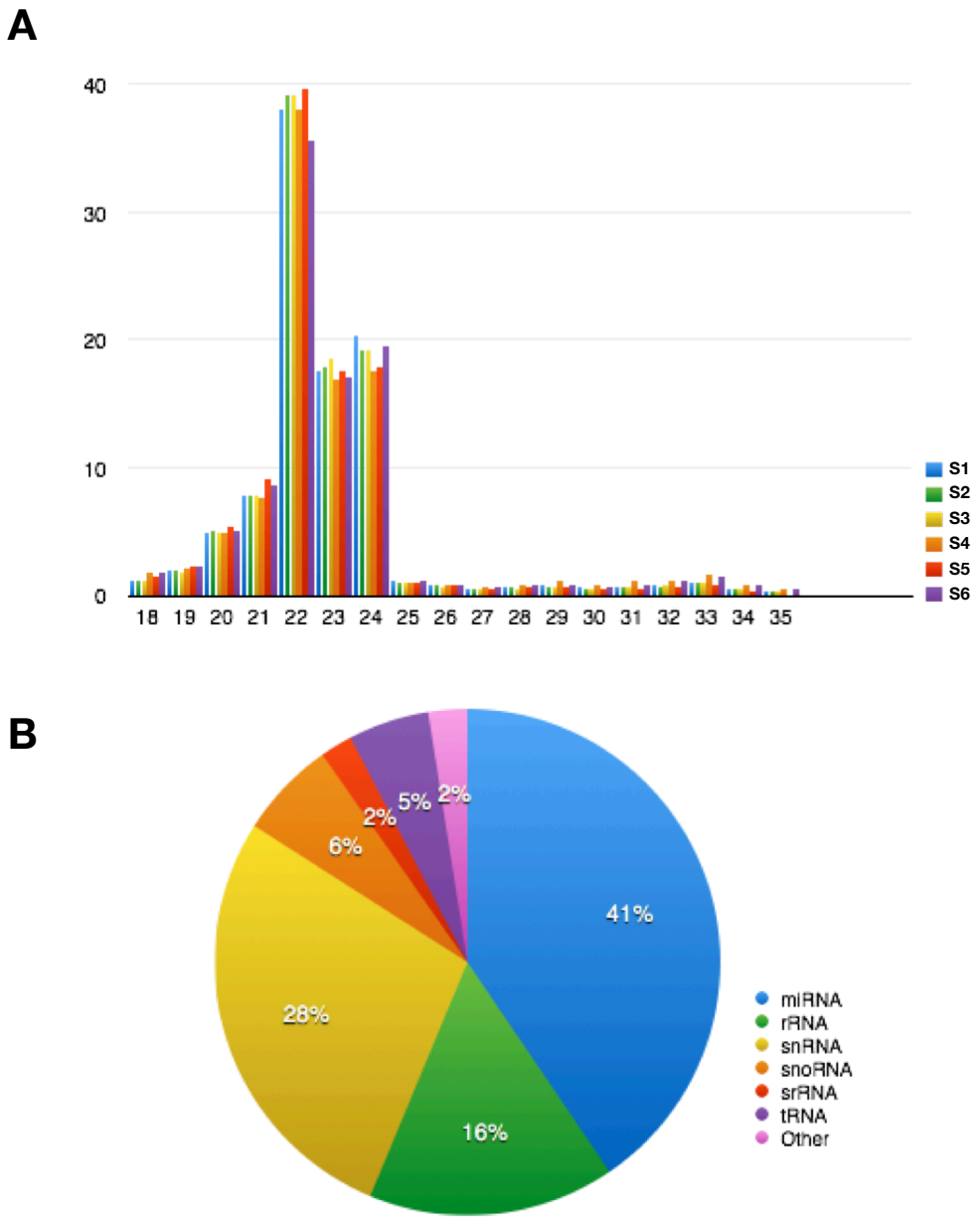

Figure 2.9 – Distribution of intracellular sRNAs sizes and types. A. Intracellular sRNAs showed a pronounced peak at 22nt, in accordance with a typical miRNA-type size distribution curve. B. The distribution of reads based on homology showed that the most predominant portion of intracellular sRNAs clean reads (41%) were similar to previously described miRNAs contained in RFAM database (source: https://rfam.xfam.org/).

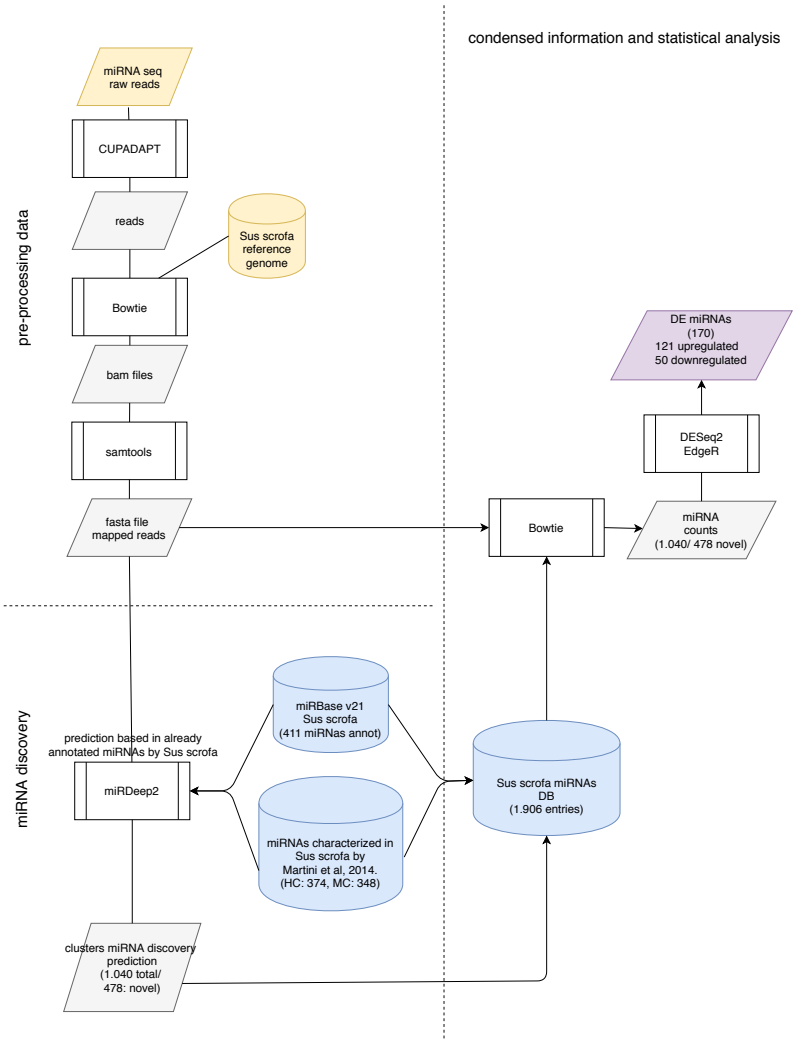

Figure 2.10 – Complete pipeline for miRNA mapping and prediction. After filtering low quality reads and trimming adapters with CutAdapt, clean sRNA-seq reads were mapped against the porcine genome with Bowtie. Intracellular sRNA samples were used as input for miRDeep2 to predict miRNAs and we kept predictions with a score of at least 5. Next, we collapsed similar predictions and obtained a total of 773 clusters (773 miRNAs), of which 478 were novel miRNAs. These predictions were included in a porcine miRNA DB (ssc-miRNA-DB) along with the 411 annotated porcine miRNAs in version 21 of miRBase and with the 722 miRNAs characterized by Martini et al. (2014). Reads from all samples were mapped against this database with Bowtie and a matrix of counts was generated in order to identify DE miRNAs. Differential expression was performed with miRNAs that had at least 50 counts across all libraries. DESeq2 and EdgeR were used for intracellular and extracellular samples and GFold was used for exosome sequences. In total were identified 170 miRNAs DE, of which 121 were up-regulated and 50 were down-regulated.

miRNA differential expression

In all sRNA samples, we detected (with at least 10 counts across all libraries) 290 from the 411 miRNAs from miRBase and 214 of the miRNAs described by Martini *et al.*, (2014) [138]. Also, 263 predicted miRNAs had more than 10 counts across all libraries. We only took into account the 491 miRNAs that had more than 50 counts across all libraries for differential expression analysis.

In total, we identified 170 DE miRNAs (121 up-regulated, 48 down-regulated and 1 ambiguous). Table 2.8 shows 10 selected up-regulated and down-regulated miRNAs detected in this study. Several homologs of these miRNAs have already been linked to bacterial infection response or immune system response in the literature and these references are listed in Table 2.8. With the exception of one ambiguous case (ssc-miR-9842-5p), whenever a miRNA was detected as DE in more than one condition (INTRA, EXTRA or EXO), the change in expression (either up- or down-regulated) was in accordance between them.

miRNA ID	DB Type	Intracellular		Extracellular		Exosomes		Play a role in other bacterial infection
		DE	LogFC	DE	LogFC	DE	GFOLD	
Novel-Chr13-miR-10	novel	Upregulated	4.121[*1]	Upregulated	9.420	Upregulated	3.659	
Novel-Chr4-miR-57	novel	Upregulated	1.218	Upregulated	12.164	Upregulated	11.582	
Novel-Chr9-miR-16	novel	Upregulated	0.768	Upregulated	13.079	Upregulated	7.947	
Novel-Chr9-dna-26	novel	Upregulated	1.448	Upregulated	11.270	Upregulated	3.592	*Pseudomonas aeruginosa*[40][*2]
mmu-mir-2143-2_6__251367_251451_-_3p_-5-424-HC	Martini *et al.*, 2014	Upregulated	2.053[*1]	Upregulated	5.899	Upregulated	2.187	
ssc-miR-1285	mirBase	NS		Upregulated	4.965	NS		*Chlamydia trachomatis*[46, 45]
ssc-miR-196b-5p	mirBase	NS		Upregulated	3.971	NS		*Mycobacterium avium*[119]
ssc-miR-212	mirBase	NS		Upregulated	8.723	NS		*Mycobacterium tuberculosis*[184]
ssc-miR-24-1-5p	mirBase	NS		Upregulated	9.742	NS		*Escherichia coli* and *Staphylococcus aureus*[156]
ssc-miR-146a-5p	mirBase	Upregulated	1.015	ND		ND		*Helicobacter pylori*[124, 116, 123] LPS stimulation[34]
ssc-miR-9842-5p	mirBase	NS		Upregulated	4.484	Downregulated	-2.497	
ssc-miR-184	mirBase	Downregulated	-0.594	ND		NS		*Chlamydia trachomatis*[47]
ssc-miR-140-3p	mirBase	NS		NS		Downregulated	-3.536	*Mycobacterium tuberculosis*[217]
ssc-miR-769-3p	mirBase	NS		Downregulated	-3.924[*1]	NS		
ssc-miR-101	mirBase	NS		Downregulated	-4.154[*1]	Downregulated	-2.986	*Helicobacter pylori*[141]
ssc-miR-107	mirBase	NS		Downregulated	-2.515[*1]	Downregulated	-2.052	Gut microbiota[209]
ssc-miR-31	mirBase	NS		Downregulated	-1.167[*1]	Downregulated	-2.004	*Helicobacter pylori*[141] T-cell activation[208]
ssc-miR-532-5p	mirBase	NS		Downregulated	-3.343	Downregulated	-3.916	LPS stimulation[34]
antisense-pn8/bta-mir-2320_6__43886629_43886722_-_NA-424-HC	Martini *et al.*, 2014	NS		Downregulated	-3.150[*1]	Downregulated	-4.440	*Actinobacillus pleuropneumoniae*[162][*3]
antisense-ssc-mir-320a*_14_6473893_6473973_-_NA;antisense-sscmir-320a*_14_6473894_6473960_-_NA-424-HC	Martini *et al.*, 2014	NS		Downregulated	-1.907[*1]	Downregulated	-2.933	*Helicobacter pylori*[158]
ssc-mir-107-shorter/ssc-isomir-107_14_106321702_106321788_-_NA-424-HC	Martini *et al.*, 2014	NS		Downregulated	-2.582[*1]	Downregulated	-2.206	Gut microbiota[209]

[*1]: Whenever LogFC from DESeq2 was not available, we provide LogFC from EdgeR;
[*2]: Reference related to **cel-miR-233-5p**, a possible homolog of novel miRNA 9-dna-26;
[*3]: *A. pleuropneumoniae* was also related to a swine cell infection.

Table 2.8 – **Selected DE miRNAs.** Information about selected up- and down-regulated miRNAs in intracellular, extracellular and exosome samples. Several homologues of these miRNAs were already described to be involved with bacterial infection in other species, and references are provided whenever we found a correlation in the same direction of expression as in this work. The only miRNA that showed ambiguous expression among conditions was ssc-miR-9842-5p, which had inverse expression between the extracellular (up-regulated) and exosome (down-regulated) samples. NS: Not significant; ND: Not detected.

Targets of DE miRNAs were enriched in genes related to redox homeostasis, translation and cytoskeleton

In order to better understand the biological functions that could be involved with the 170 DE miRNAs, we performed different analyses to predict their potential targets. In this sense, all DE miRNAs had at least one DE gene as a predicted target. The complete pipeline used for target prediction in this study is provided in Figure 2.11.

We predicted a total of 79,276 interaction pairs between miRNAs and mRNAs. In this way, based only on the predictions of the software used here, a miRNA could potentially target, on average, 465 genes in the entire porcine genome. However, we only considered as targets the mRNAs that were detected as DE in this study, which significantly decreased our list to a total of 4,287 interaction pairs. We chose to focus on anticorrelations between miRNA and mRNA expression (and kept 1,939 interaction pairs), as the main mode of action of miRNAs is a destabilization of mRNAs [16] and the fact that most of experimental validations in the literature are related to interaction pairs with inversed regulation. In this context, a permissive interaction is generally described as one that occurs between a down-regulated miRNA and an up-regulated mRNA, while a repressive interaction is one in which the miRNA is up-regulated with consequent down-regulation of the target mRNA [138] (Figure 2.12A). In our results, permissive interactions represented 267 genes and 50 miRNAs, while repressive interactions occurred between 425 genes and 121 miRNAs, accounting for 598 permissive and

1341 repressive interactions between DE mRNAs and miRNAs.

We performed GO analyses to compare the DE genes and targets of DE miRNAs in order to investigate whether their functions could be correlated (Figure 2.12B and Figure 2.13). Interestingly, we were able to identify a correlation between the enriched terms in miRNA targets with some of the GO terms detected in the mRNA up-regulated or down-regulated GO results (Figure 2.12C). Target genes from permissive interactions were associated with ribosome/translation and oxidation-reduction activity, whereas target genes from the repressive interactions had enriched terms related to cytoskeleton and ciliary function (Figure 2.12C). It is important to highlight the relevance of the miRNAs found in exosome-like vesicles and in extracellular samples in the identification of GO enriched terms of the miRNA targets, since only a small part of the permissive and repressive interactions involved intracellular miRNAs. Complete GO enrichemnt of the miRNA targets is found in Table 2.9.

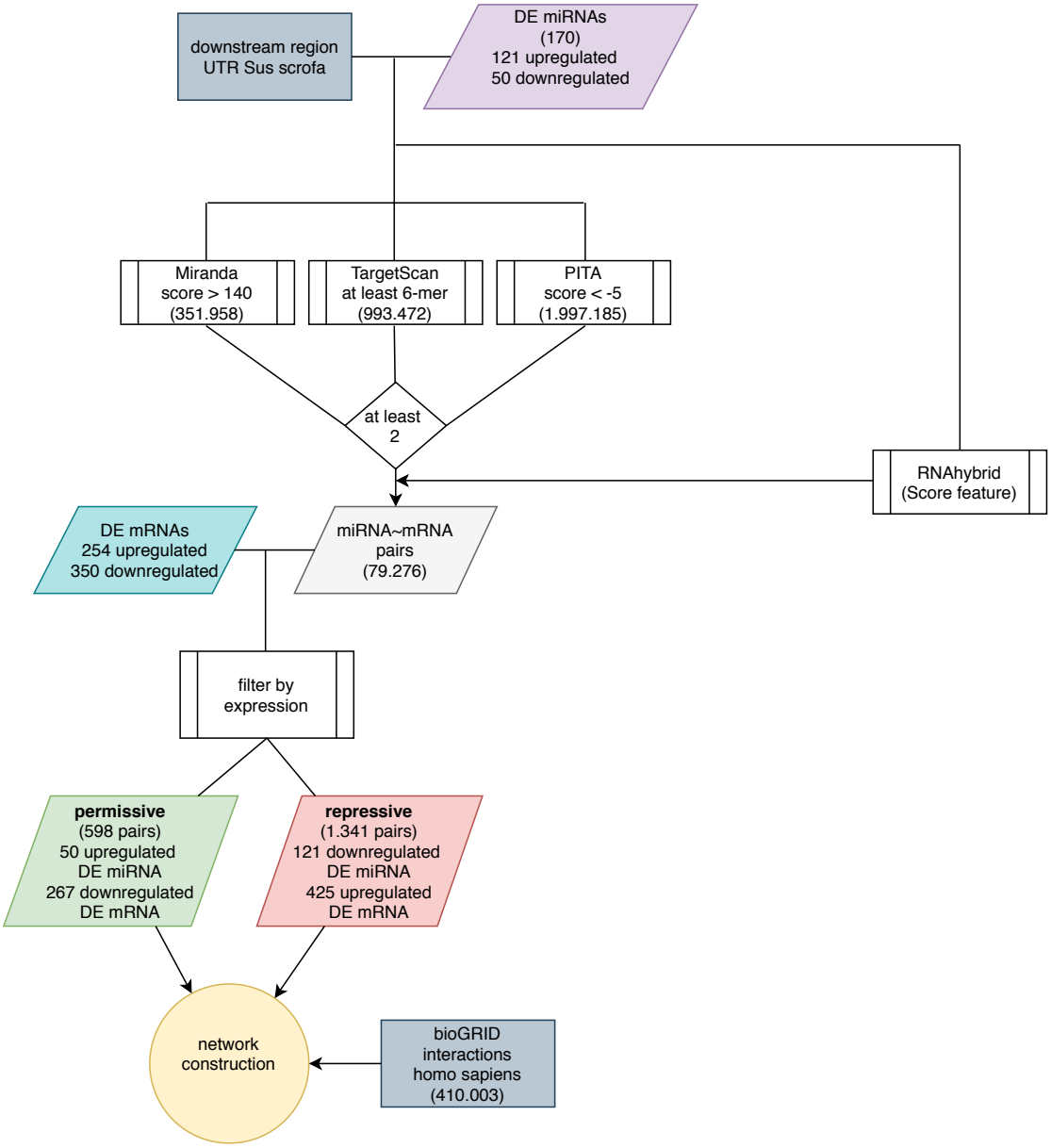

Figure 2.11 – Complete pipeline for miRNA target prediction. DE miRNAs were used as input to detect putative interactions with the 3'UTRs of Ensembl transcripts in the porcine genome. TargetScan, miRanda and Pita were used to detect target pairs and RNAhybrid was used to validate the hybridization of a target pair. The following thresholds were used: score in miRanda > 140, DDG from Pita <-5, score in RNAhybrid <-15 and prediction in TargetScan of at least 6mers. Only targets predicted with a good score for at least two distinct tools were kept. After these, we selected from the list only the target genes that were detected as DE in this study, and subsequentially we only considered target pairs of miRNA-mRNA that had inversed fold change expression. These pairs with inversed correlation (permissive and repressive) were used for the network reconstruction in Cytoscape along with information about interactions from bioGRID.

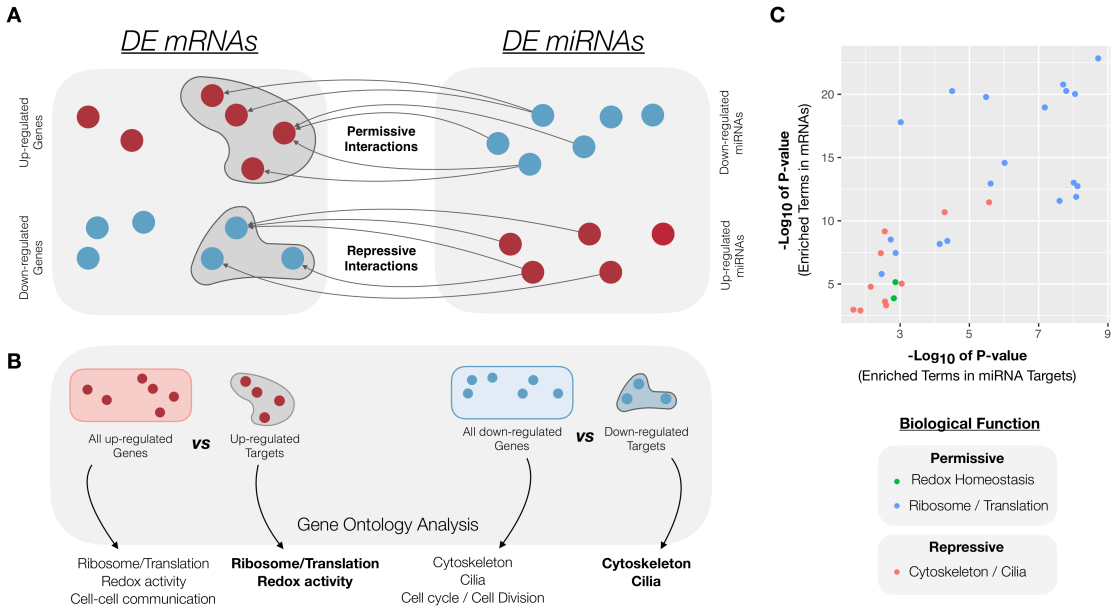

Figure 2.12 – **Correlation of GO terms between DE mRNAs and targets of DE miRNAs. A** A permissive interaction occurs between a down-regulated miRNA (depicted in blue) and an up-regulated mRNA (depicted in red), whereas a repressive interaction is one in which the miRNA is up-regulated (red) and its target mRNA is down-regulated (blue). **B.** We performed GO enrichment analyses with the complete up- and down-regulated lists of mRNAs and also with the subset of miRNA targets among each of these lists. **C.** A correlation of some of the GO terms from DE mRNAs and targets of DE miRNAs was detected, indicating that these functions might also be regulated by miRNAs.

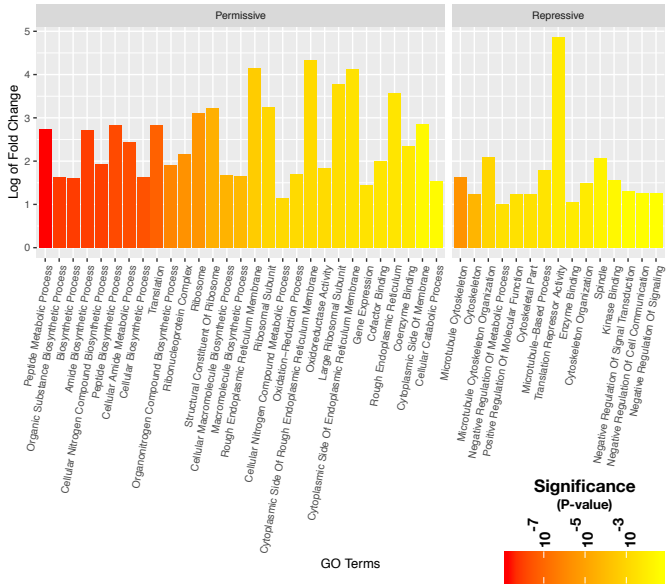

Figure 2.13 – Complete GO enrichment analysis for miRNA targets. Target genes from permissive interactions were enriched in terms related to ribosome/translation and oxidation-reduction activity, whereas target genes from the repressive interactions were associated to cytoskeleton and ciliary function.

GO	Term	FC	P-value	Over/Underrepresented	Type	DE
GO:0034641	Cellular Nitrogen Compound Metabolic Process	2.22	1.35E-03	+	BP	Permissive
GO:0010467	Gene Expression	2.70	3.43E-03	+	BP	Permissive
GO:0044248	Cellular Catabolic Process	2.89	3.16E-02	+	BP	Permissive
GO:0009058	Biosynthetic Process	3.03	8.26E-09	+	BP	Permissive
GO:0044249	Cellular Biosynthetic Process	3.07	2.49E-08	+	BP	Permissive
GO:1901576	Organic Substance Biosynthetic Process	3.08	7.57E-09	+	BP	Permissive
GO:0009059	Macromolecule Biosynthetic Process	3.14	7.21E-05	+	BP	Permissive
GO:0034645	Cellular Macromolecule Biosynthetic Process	3.21	4.33E-05	+	BP	Permissive
GO:0055114	Oxidation-Reduction Process	3.27	1.37E-03	+	BP	Permissive
GO:0016491	Oxidoreductase Activity	3.56	1.52E-03	+	MF	Permissive
GO:1901566	Organonitrogen Compound Biosynthetic Process	3.76	9.61E-07	+	BP	Permissive
GO:0044271	Cellular Nitrogen Compound Biosynthetic Process	3.83	9.69E-09	+	BP	Permissive
GO:0048037	Cofactor Binding	4.00	3.90E-03	+	MF	Permissive
GO:1990904	Ribonucleoprotein Complex	4.47	2.44E-06	+	CC	Permissive
GO:0050662	Coenzyme Binding	5.09	5.33E-03	+	MF	Permissive
GO:0043603	Cellular Amide Metabolic Process	5.46	1.96E-08	+	BP	Permissive
GO:0043604	Amide Biosynthetic Process	6.57	8.91E-09	+	BP	Permissive
GO:0006518	Peptide Metabolic Process	6.71	1.91E-09	+	BP	Permissive
GO:0006412	Translation	7.06	6.64E-08	+	BP	Permissive
GO:0043043	Peptide Biosynthetic Process	7.17	1.60E-08	+	BP	Permissive
GO:0098562	Cytoplasmic Side Of Membrane	7.20	2.60E-02	+	CC	Permissive
GO:0005840	Ribosome	8.56	3.30E-06	+	CC	Permissive
GO:0003735	Structural Constituent Of Ribosome	9.30	3.15E-05	+	MF	Permissive
GO:0044391	Ribosomal Subunit	9.54	9.61E-04	+	CC	Permissive
GO:0005791	Rough Endoplasmic Reticulum	11.96	4.08E-03	+	CC	Permissive
GO:0015934	Large Ribosomal Subunit	13.63	1.87E-03	+	CC	Permissive
GO:0098554	Cytoplasmic Side Of Endoplasmic Reticulum Membrane	17.33	3.04E-03	+	CC	Permissive
GO:0030867	Rough Endoplasmic Reticulum Membrane	17.76	3.85E-04	+	CC	Permissive
GO:0098556	Cytoplasmic Side Of Rough Endoplasmic Reticulum Membrane	20.10	1.44E-03	+	CC	Permissive
GO:0009892	Negative Regulation Of Metabolic Process	2.00	2.53E-03	+	BP	Repressive
GO:0019899	Enzyme Binding	2.06	6.92E-03	+	MF	Repressive
GO:0044430	Cytoskeletal Part	2.34	2.78E-03	+	CC	Repressive
GO:0044093	Positive Regulation Of Molecular Function	2.36	2.74E-03	+	BP	Repressive
GO:0005856	Cytoskeleton	2.37	5.19E-05	+	CC	Repressive
GO:0023057	Negative Regulation Of Signaling	2.39	4.49E-02	+	BP	Repressive
GO:0010648	Negative Regulation Of Cell Communication	2.40	4.18E-02	+	BP	Repressive
GO:0009968	Negative Regulation Of Signal Transduction	2.47	3.77E-02	+	BP	Repressive
GO:0007010	Cytoskeleton Organization	2.83	7.06E-03	+	BP	Repressive
GO:0019900	Kinase Binding	2.93	2.24E-02	+	MF	Repressive
GO:0015630	Microtubule Cytoskeleton	3.08	2.70E-06	+	CC	Repressive
GO:0007017	Microtubule-Based Process	3.48	3.61E-03	+	BP	Repressive
GO:0005819	Spindle	4.18	1.40E-02	+	CC	Repressive
GO:0000226	Microtubule Cytoskeleton Organization	4.29	9.00E-04	+	BP	Repressive
GO:0030371	Translation Repressor Activity	29.21	6.25E-03	+	MF	Repressive

Table 2.9 – **GO enrichment DE miRNAs.**

Regulatory network reconstruction and analysis

In order to gain a broader view of the host response to the presence of *M. hyopneumoniae*, we built a general regulatory network by integrating mRNA gene expression with predicted miRNA-mRNA interactions collected and analyzed in this work (Figure 2.2). We also included information about physical and genetic validated interactions from the BioGRID v3.4 database [189, 29]. As previously mentioned, at this point we only took into account interaction pairs that were either repressive or permissive. The global interaction network was composed by 774 nodes and 1965 arcs and our objective was to identify miRNA-mRNA expression patterns. We performed a functional analysis with the use of CLUEGO [19] and we also detected within the permissive interactions the enrichment of several processes related to immune response and inflammation (Figure 2.14).

Furthermore, we created two separate regulatory networks, one related to cytoskeleton and cilia (repressive interactions, Figure 2.15A) and one related to redox homeostasis (permissive interactions, Figure 2.15B) to better refine and understand the possible co-regulation of several of these targets. The networks show a high level of connectivity and we were able to detect miRNAs that interacted only with NRF2 activated targets (such as ssc-miR-31) (Table 2.10). Furthermore, we were able to detect miRNAs that seemed to regulate more generally redox homeostasis genes and also miRNAs whose targets came from sets of genes with distinct functions (such as glycolysis, immune system defense, ribosomes, among others), indicating that this response might be related to several other important functions within the cell. We believe that this network can be a powerful tool for analyzing the influence of *M. hyopneumoniae* on the host gene expression in future studies.

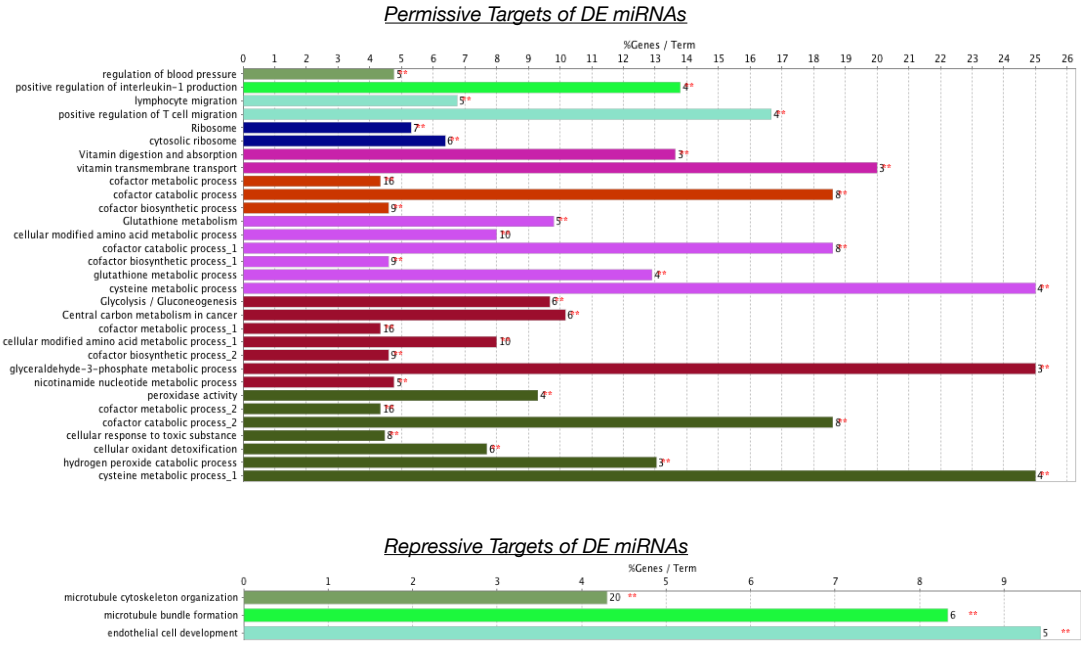

Figure 2.14 – CluеGO analysis of the repressive and permissive pairs. In the repressive interactions we detected the enrichment of terms related to cytoskeleton, while in the permissive interactions, besides terms related to translation and oxidation-reduction activity, there was an enrichment of several processes related to immune response and inflammation.This analysis was performed with the complete list of target genes in either repressive or permissive interaction networks with a threshold for p-value of 0.01.

Permissive interactions were enriched in genes regulated by NRF2

Analyzing more in detail the permissive interactions between mRNAs and miRNAs, we identified several miRNAs targeting NRF2 regulated genes (Table 2.10). *GGT1*, for instance, is predicted to be targeted by 7 different miRNAs, while ssc-mir-31 is predicted to target 4 different genes induced by NRF2 (*AKR1C4, AKR1CL1, GGT1*, and *TXNRD1*).

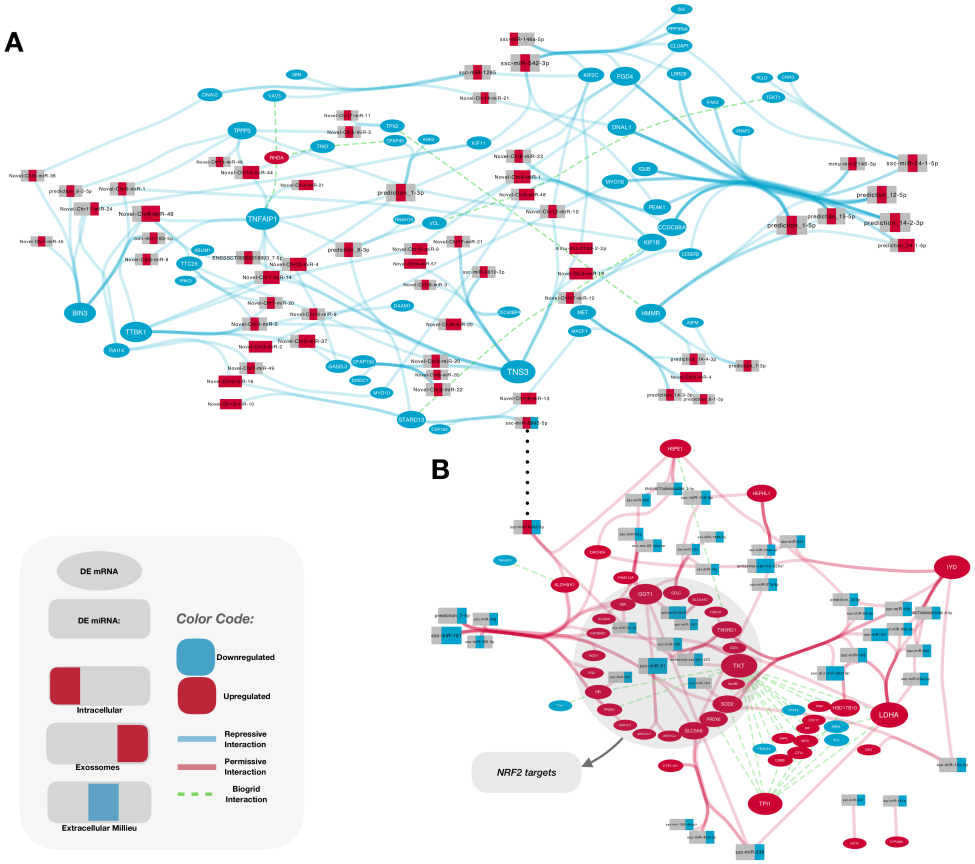

Figure 2.15 – **Interaction networks of DE miRNAs and genes involved with cytoskeleton/cilia and redox homeostasis.** Rectangles depict miRNAs and ellipses represent genes. Blue ellipses/rectangles indicate that the gene/miRNA was down-regulated and red indicate gene/miRNAs detected as up-regulated. BioGrid interactions are seen as green dashed arcs. These represent both physical and genetic interactions between either genes or proteins from such genes. **A.** Cytoskeleton/cilia interaction network. We detected many repressive interactions involving genes related to cytoskeleton and ciliary function (blue arcs). **B.** Redox homeostasis interaction network. Permissive interactions of genes related to redox homeostasis are shown as red arcs. The great majority of the miRNAs targeting these genes are DE only in exosomes (rectangles partially colored in blue, far right). NRF2 targets: genes described in other species to be activated by NRF2 are highlighted in grey.

Upregulated Gene ID	Downregulated miRNA	miRNA Expression		
		Intracellular	Extracellular	Exosome
AKR1C1	ssc-miR-101	NS	Downregulated	Downregulated
AKR1C4	ssc-miR-31	NS	Downregulated	Downregulated
AKR1CL1	ssc-miR-31	NS	Downregulated	Downregulated
BLVRB	prediction_13_40037950_40038026_-_3p-353-MC	NS	ND	Downregulated
EGR1	antisense-ssc-mir-423_12_44150500_44150579_-_NA-353-MC	NS	NS	Downregulated
GCLC	ssc-miR-148b-3p	NS	NS	Downregulated
	ssc-miR-19a	NS	NS	Downregulated
	ssc-miR-19b	NS	NS	Downregulated
GGT1	ssc-miR-338	NS	ND	Downregulated
	ssc-miR-140-3p	NS	NS	Downregulated
	ssc-mir-92-shorter/ssc-isomir-92_X_108178486_108178565_-_NA;ssc-mir-92-shorter/ssc-isomir-92_X_108212408_108212487_-_NA-424-HC	NS	NS	Downregulated
	ssc-miR-92a	NS	NS	Downregulated
	prediction_7_91732076_91732132_+_5p-353-MC	NS	NS	Downregulated
	ssc-miR-31	NS	Downregulated	Downregulated
	antisense-pn8/bta-mir-2320_6_43886629_43886722_-_NA-424-HC	NS	Downregulated	Downregulated
GSR	ssc-miR-101	NS	Downregulated	Downregulated
	ssc-miR-22-3p	NS	NS	Downregulated
HMOX1	ssc-miR-182	NS	NS	Downregulated
HSP90AB1	ssc-miR-101	NS	Downregulated	Downregulated
NQO1	ssc-miR-34a	NS	NS	Downregulated
PGD	ssc-miR-769-3p	NS	Downregulated	NS
PIR	ssc-miR-378	NS	NS	Downregulated
PRDX1	ssc-miR-101	NS	Downregulated	Downregulated
PRDX6	ssc-miR-10a-5p	NS	NS	Downregulated
	ssc-miR-10b	NS	NS	Downregulated
	ssc-miR-338	NS	ND	Downregulated
SLC5A10	ssc-miR-1307	NS	ND	Downregulated
	antisense-pn8/bta-mir-2320_6_43886629_43886722_-_NA-424-HC	NS	Downregulated	Downregulated
SLC5A6	ssc-miR-10b	NS	NS	Downregulated
	ssc-miR-338	NS	ND	Downregulated
	ssc-miR-423-3p	NS	ND	Downregulated
	ssc-miR-27b-3p	NS	NS	Downregulated
	prediction_7_91732076_91732132_+_5p-353-MC	NS	NS	Downregulated
	ssc-miR-340	NS	ND	Downregulated
SOD2	prediction_7_91732076_91732132_+_5p-353-MC	NS	NS	Downregulated
	ssc-miR-184	Downregulated	ND	ND
	ssc-miR-28-3p	NS	NS	Downregulated
TKT	antisense-ssc-mir-423_12_44150500_44150579_-_NA-353-MC	NS	NS	Downregulated
TXNRD1	ssc-miR-103	NS	NS	Downregulated
	ssc-miR-182	NS	NS	Downregulated
	ssc-mir-107-shorter/ssc-isomir-107_14_106321702_106321788_-_NA-424-HC	NS	Downregulated	Downregulated
	ssc-miR-107	NS	Downregulated	Downregulated
	ssc-miR-31	NS	Downregulated	Downregulated
UGT1A6	ssc-miR-28-3p	NS	NS	Downregulated
	antisense-ssc-mir-103_16_52667657_52667738_+_NA-424-HC	NS	NS	Downregulated
	ssc-miR-769-3p	NS	Downregulated	NS

Table 2.10 – **DE genes putatively activated by NRF2 predicted to be targets of DE miRNAs.** Expression of miRNAs is reported for intracellular, extracellular and exosome-like vesicles. NS: Not significant; ND: Not detected.

It was previously reported that *NRF2* can be regulated independently of KEAP1 by miRNAs in human breast cancer cells [211]. Our miRNA target analysis predicted *NRF2* as target of three miRNAs down-regulated in exosome-like vesicles: ssc-mir-340, ssc-mir-19a, and ssc-mir-19b. The first was also predicted to target the transporter gene *SLC6A6*, while ssc-mir-19a and ssc-mir-19b were predicted to target *GCLC*, necessary for glutathione synthesis. In addition, miR-340 has been previously identified negatively regulating *NRF2* expression in human [181]. Conversely, miR-19a and miR-19b have been reported to be down-regulated in the presence of hydrogen peroxide in rats [205, 80], linking these miRNAs to a response to oxidative stress. In the literature, other miRNAs were also shown to negatively regulate the expression of *NRF2*: miR-28 [211], miR-101 [65], miR-92a [122], miR-27b [210], and miR-34a [5]. All of them were down-regulated in our data and were predicted by this study to regulate NRF2

activated genes (Table 2.10). In this way, it seems that during *M. hyopneumoniae* infection, there is a global change in gene expression in an attempt to activate antioxidant genes, in association with the down-regulation of miRNAs that negatively regulate these genes.

Repressive interactions were related to cytoskeleton and cilia

As concerns the repressive interactions of miRNAs and mRNAs, we identified within the targets of up-regulated miRNAs an enrichment of genes related to cytoskeleton and cilia. For instance, the gene *TNS3*, related to cytoskeleton, is predicted as target of 16 up-regulated miRNAs. In addition, several genes encoding for dyneins were predicted as targets of up-regulated miRNAs, with *DNAL1* being target of 7 different miRNAs. Mutations and down-regulation of genes coding ciliary proteins, especially dyneins, were related to primary ciliary dyskinesia [57, 146, 191, 177, 83], but had not yet been reported in *M. hyopneumoniae* infection. More investigations on this matter should be carried out, since these results seem to be related to one of the main adverse effects caused by this bacterium: ciliostasis.

Most DE miRNAs were detected in exosome-like vesicles

Analyzing the up-regulated genes, we identified that besides antioxidant response and protein synthesis, there is an enrichment in GO terms related to cell-cell communication (Figure 2.4). The up-regulation of cell-cell communication is noteworthy considering that most DE miRNAs were detected only in the extracellular and exosome-like samples. Moreover, most of the down-regulated miRNAs of exosome-like vesicles were DE only in these vesicles, indicating that the cells might be selecting specific populations of miRNAs to be packaged into these structures, which could be seen as a specific message they are trying to communicate to other cells. In the same direction, the increase in the number of proteins secreted in vesicles was reported in NPTr cells infected with *M. hyopneumoniae*[222], indicating that the communication via exosome-like vesicles is important for swine cells during infection. This corroborates our data, suggesting that *M. hyopneumoniae* induces modification on the protein and RNA composition of NPTr-released vesicles and is likely important in the cross-talks between epithelial cells.
The relationship between cell-cell communication and DE miRNAs released by infected cells becomes especially interesting when considering genes related to the NRF2 pathway: we identified DE miRNAs predicted to regulate such genes only in extracellular and exosome-like vesicle samples, with no difference of expression in intracellular samples (Table 2.10). This might suggest the existence of a mechanism in which infected cells send signals to neighboring cells in order to prevent the repression of these genes or degradation of their RNA products. This becomes even more important due to the fact that exosomal miRNAs were already shown to regulate the inflammatory response in receptor cells from mice [4]. While miRNAs identified in the extracellular total samples might contain sRNAs that reflect degradation remnants due to the presence of RNAses and mRNAs in the extracellular environment, we believe that this is less likely to happen in exosomes, since these vesicles have a membrane that

protects their content from extracellular degradation. As exosomes have an important role in cell communication, we should consider the difference in miRNA expression in these vesicles as relevant and speculate on how they might interfere in the gene regulation of neighboring cells. In addition, as exosomes may also affect other cell types, they may contain miRNAs that target genes that were not detected as DE in our samples.

2.4 Conclusion.

M. hyopneumoniae is considered a pathogen with huge negative impact on swine production. However, apart from studies related to adherence to the host cells, little is known about the relationship between this pathogen and the swine host. In this work, we analyzed the changes that *M. hyopneumoniae* induced in gene and miRNA expression in tracheal epithelial cells. As far as we know, this is the first study to establish a link between gene expression of the swine cells and the most deleterious pathogenic effects of *M. hyopneumoniae*, namely its cytotoxic epithelial damage (possibly via hydrogen peroxide production) and induced ciliostasis. However, we are aware that infection with this species involves a large component of the immune system, which is even said to be responsible for the major tissue damage related to the infection. Thus, the results found here reflect the effects of *M. hyopneumoniae* on epithelial cells, while the general picture of the respiratory tract might present different responses to the infection. In this way, it is important to observe that besides the hydrogen peroxide production by this bacterium, other factors can trigger the activation of the cell antioxidant defense during infection. *M. hyopneumoniae* infection is characterized by the infiltration of a large number of leukocytes in the lung tissue, which produce reactive oxygen and nitrogen species, causing tissue damage [153, 21]. Accordingly, systemic infections with *M. hyopneumoniae* have a great potential of causing oxidative stress, so that the transcription of genes activated by NRF2 seems to be important to fight the infection and maintain cellular homeostasis.
In conclusion, we conducted the first study that analyzes mRNA and miRNA differential expression by NGS in epithelial tracheal cells infected with *M. hyopneumoniae*. Our results bring new insights into the interaction between this bacterium and swine epithelial cells, notably the host cellular response through the activation of genes related to antioxidant response and the repression of cytoskeleton and ciliary genes (possibly related to ciliostasis), and open several perspectives related to the understanding of the pathogenicity of this bacterial species. The work presented in this Chapter was done in a collaborative project.

Chapter 3

BrumiR algorithmn

Contents

3.1 Introduction. **65**

3.2 Definitions. **68**

3.3 Implementation. **68**

3.3.1 Building a de Bruijn graph for sRNA-seq data. 68

3.3.2 Removing sequencing errors from the unipath sRNA-seq graph. 72

3.3.3 An expressed mature miRNA has uniform coverage. 72

3.3.4 miRNAs and other sequences are captured in single connected components. 72

3.3.5 BrumiR classifies low abundance non-linear topologies as sequencing artefacts. 73

3.3.6 Re-assembling unipaths within each CC. 73

3.3.7 Re-clustering potential miRNAs. 73

3.3.8 Identifying other expressed RNA sequences. 76

3.4 BrumiR algorithm: from miRNA reads to a de Bruijn graph. . . . **78**

3.5 Results: BrumiR achieves the highest accuracy on simulated data. **81**

3.6 Conclusion. **86**

3.1 Introduction.

Nowadays, a common experimental practice is to identify miRNAs and their expression patterns using next generation sequencing technologies (NGS) [151] which represent a powerful

tool to obtain genome-wide expression levels of sRNAs. Commonly, NGS experiments are able to generate more than 20 million sRNA-seq reads, thus promoting the development of algorithms to transform and process such data into biological information [32].

As mentioned earlier, there are two computational strategies for the discovery of miRNAs: 1) genome-based approaches that rely on the mapping of the sRNA-seq reads to a reference genome and subsequent evaluation of the sequences generating the characteristic hairpin structure of miRNA precursors [23]; 2) machine-learning approaches which rely on the biogenesis features extracted from the knowledge on miRNA sequences available in databases such as miRBase [97] and on the analysis of the duplex structure of miRNAs [195]. Genome-based methods, that have been updated at the pace of the evolving NGS technologies, are the most widely used tools in this field, and their results have populated the public miRNA repositories [32]. Such methods are the natural choice for the study of model species with high quality reference genomes available. However, it has been shown that most of the genome-based tools struggle with a high rate of false positive predictions [23]. Additionally, a critical step of such tools is the use of genome aligners [103, 115] to map the sRNA-seq reads to the reference genome. Mapping short (<30 nt) and very similar sequences to a large, complex, and repetitive reference genome is however a difficult and error-prone task [221]. Genome-based methods are thus highly sensitive to the aligner selected as well as to the parameters employed and the thresholds chosen (*e.g.* number of mismatches allowed) in order to discard mapping artefacts generated from sequencing errors [118]. Furthermore, despite all the advancements in the sequencing technologies and *de novo* assembly methods, few complete genomes are available today, which is a recurring problem that researchers working on non-model species face [1]. The lack of a high quality reference genome thus reduces the possibilities for discovering novel miRNAs [195]. Genome-based methods such as miRDeep [61], miRDeep2 [63], and miR-PREFeR [110] are included in this group.

On the other hand, new methods such as miReader [84], MirPlex [135], and mirnovo [195], in particular using machine-learning approaches, were specifically developed as an alternative to discover miRNAs in species without a reference genome. In the case of mirnovo, the initial step involves the clustering of the sRNA-seq reads performing an all-vs-all read comparison that is followed by a subsequent classification of the clusters into putative miRNAs using pre-trained models. The performance obtained by such methods on well-annotated species is comparable to those achieved by genome-based methods [23]. However, relying exclusively on annotated miRNAs for training machine learning models may introduce a bias towards the identification of well-characterized miRNAs over species-specific ones [32]. Nonetheless, machine learning methods have demonstrated that it is possible to discover miRNAs using only the sequence information present in the sRNA-seq experiment [195].

There remains however a need to go further in the development of algorithms for finding novel miRNAs in non-model species using only the sequence information. With this purpose in mind, the adoption of a special type of graphs called *de Bruijn* graphs may be considered. This is a widely used approach for the *de novo* reconstruction of genome or transcriptome sequences

[39]. It therefore appears to be a plausible option for organizing, clustering and assembling the sequence information present in sRNA-seq experiments. However, accommodating the de Bruijn graph approach for the discovery of miRNAs involves the development of new methods to address the specific characteristics of sRNA-seq data. Indeed, mature miRNA sequences are short (18-24 nt), thus limiting the overlap length for building a de Bruijn graph which in turn impacts the global topology by inducing tangled graph structures. Moreover, miRNAs captured in a sRNA-seq experiment have variable expression, from low (few reads) to highly expressed (thousands of reads), which may induce spurious graph connections that should be removed in order to isolate and detect both types of miRNAs. Finally, the sequencing errors present in sRNA-seq data further induce spurious connections and are harder to detect as compared to genomic data due to the variable expression and the shorter lengths of the miRNAs. Overall, using a de Bruijn graph to analyze sRNA-seq data and extract information from such data seems thus counterintuitive as mature miRNAs are captured full-length by the current NGS technologies. However, a de Bruijn graph has several interesting properties for the discovery of miRNAs, mainly due to the fact that it encodes all the sRNA-seq sequence information at once in a compact and connected representation (graph), without the need to perform an all-vs-all read comparison or mapping to a reference.

In this chapter, we present BrumiR, a *de novo* algorithm based on a de Bruijn graph approach that is able to identify miRNAs directly and exclusively from sRNA-seq data. Unlike other state-of-the-art algorithms, BrumiR does not rely on a reference genome, on the availability of close phylogenetic species, or on conserved sequence information. Instead, BrumiR starts from a de Bruijn graph encoding all the reads and is able to directly identify putative miRNAs on the generated graph. BrumiR also removes sequencing errors and navigates inside the graph detecting putative miRNAs by considering several miRNA biogenesis properties (such as expression, length, topology in the graph). Along with miRNA discovery, BrumiR can also assemble and identify other types of small and long non-coding RNAs expressed within the sequencing data. Finally, when a reference genome is available, BrumiR provides a new mapping tool (BrumiR2reference) that performs an exhaustive search to identify and validate the precursor sequences. We benchmarked BrumiR on animal and plant species using simulated datasets. To this purpose, we developed miRsim, a tool for simulating sRNA-seq reads from mature miRNA sequences (mirBase). The benchmark results show that BrumiR is very sensitive, besides being the fastest tool, and its predictions were supported by the characteristic hairpin structure of miRNAs.

This chapter is composed as follows. In Section 3.2, I introduce some definitions. Then, in Section 3.3, I explain how we can use the de Bruijn graph for encoding sRNA-seq data, and present the different steps of BrumiR to remove noise and capture the miRNA candidates. I also show the main complexities of the de Bruijn graph using miRNA reads. In Section 3.4, I present BrumiR and in Section 3.5, I show a benchmark of BrumiR on simulated data using the miRsim tool.

3.2 Definitions.

The de Bruijn graph is constructed by breaking the reads into words of a fixed length k which are called *k-mers* (Figure 3.1A), each distinct k-mer representing a node of the graph (Figure 3.1B) while the arcs connect two nodes when the suffix of length $k-1$ of the label of one node is equal to the prefix of length $k-1$ of the other (Figure 3.1B1). The arc is then labeled by this common suffix-prefix of length $k-1$ (Figure 3.1B). For instance, if k is 5 and the vertices are labeled TGAGG and GAGGT then the arc between the first vertex and the second will be labeled GAGG (Figure 3.1B1).

The bi-directed de Bruijn graph of order k is a natural way of representing the double-stranded nature of overlaps between DNA sequences (Figure 3.1B2). In practice, two k-mers that are reverse-complements of each other are represented by a single node by selecting arbitrarily the one with smallest lexicographical order.

A sequence of arcs $e_1, ..., e_n$ in a bi-directed de Bruijn graph is a walk. A walk spells its corresponding sequence. Intuitively, an arc implies that the strings of the two nodes can be combined, using the $k-1$ overlapping bases, into a bigger sequence. For example the walk e_1, e_2, e_3, e_4 in Figure 3.1B2 spells the sequence TGAGGTAG. A walk from e_1 to e_n is a unipath if it is a path (*i.e.* does not repeat nodes) such that all the nodes have in and out-degree equal to 1 (Figure 3.1C, green nodes). A unipath is maximal if it cannot be extended in either direction. In the compacted bi-directed de Bruijn graph, every maximal unipath and its complement is replaced by a single vertex (Figure 3.1C, green nodes). Formally, the unipath graph is created by compressing the nodes into maximal unipaths (Figure 3.1C). The genome sequence is reconstructed by visiting every arc of the unipath graph exactly once (allowing the revisit of nodes), which corresponds to a Eulerian path [161] (Figure 3.1D).

In summary, the main steps to encode and build genome sequences using a de Bruijn graph approach, involves splitting the reads into k-mers, and then using the overlap of k-1 bases to build the compressed bi-directed node-centric de Bruijn graph, followed by the determination of an eulerian path to reconstruct the genomic sequence (Figure 3.1F).

BrumiR employs the Bcalm2 program to build the compressed bi-directed node-centric de Bruijn graph of order k from a set of sRNA-seq reads, plus additional algorithms described in the following sections to assemble without a reference genome a set of mature miRNA sequences.

3.3 Implementation.

3.3.1 Building a de Bruijn graph for sRNA-seq data.

BrumiR starts by building a compact de Bruijn graph from the sRNA-seq reads given as input. De Bruijn graphs are a widely used approach in the genome assembly problem [39]. BrumiR uses this graph to organize, detect, and exploit the sequence information of sRNA-

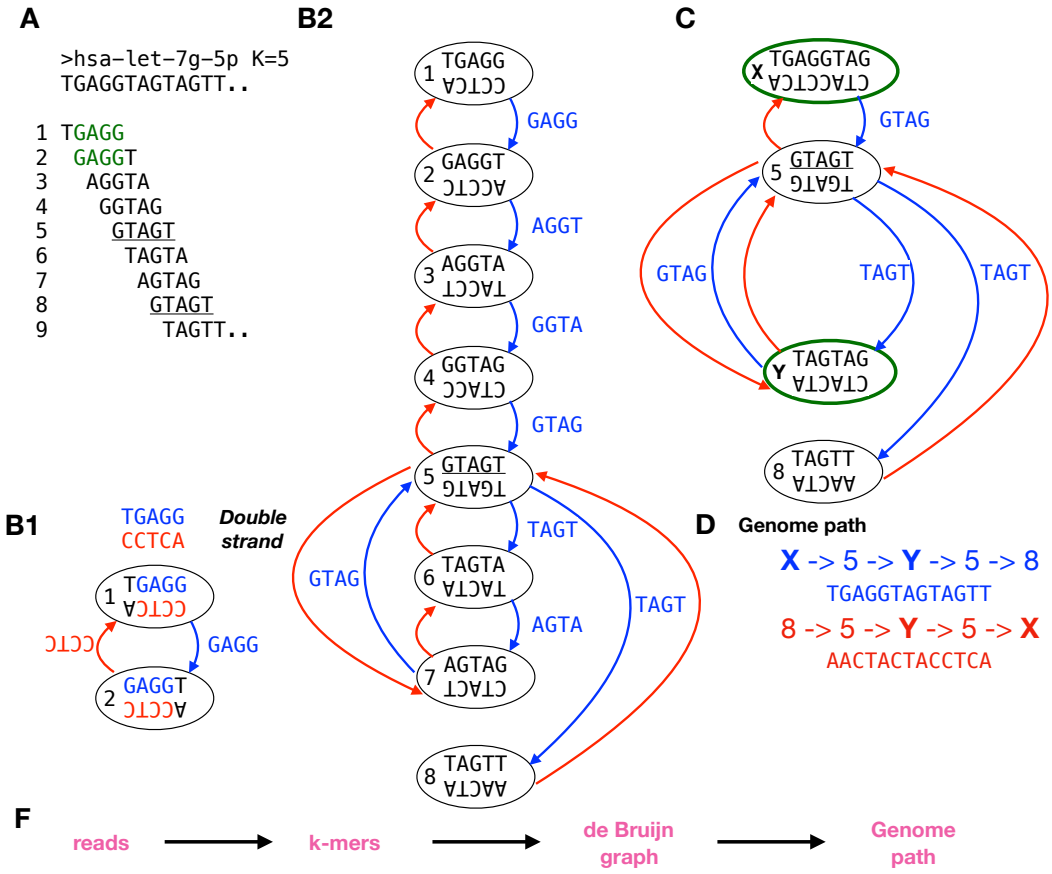

Figure 3.1 – From reads to a de Bruijn graph and from a de Bruijn graph to a genome. A) Reads are split into consecutive k-mers of a fixed length (here k=5). B) The distinct k-mers represent the nodes and an arc between two nodes is added if the k-1 suffix of one node is equal to the k-1 prefix of the other node; therefore, an arc represents an exact overlap of k-1 bases (B1). Due to the complementarity of DNA/RNA sequences, the resulting de Bruijn graph is bi-directed (B2), with blue and red arcs representing the forward and reverse strands, respectively. C) The compression of simple paths, which are composed of nodes with one in and one out arc (green nodes), leads to the so-called unipath graph. D) Paths generating the original genome sequence for both strands. F) Summary of the main steps to encode and build genome sequences using a de Bruijn graph approach.

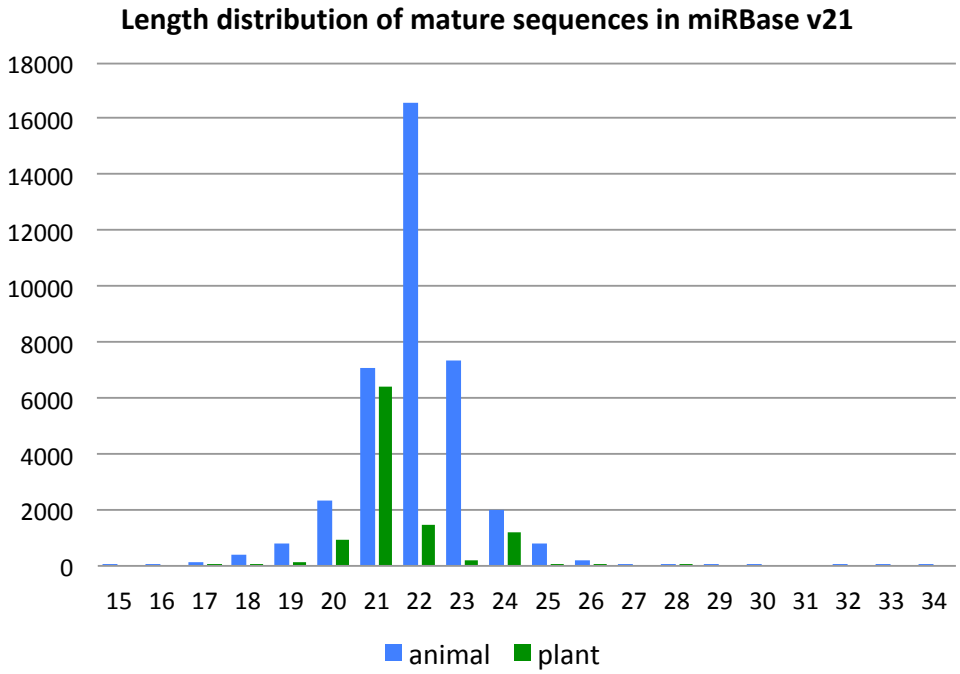

Figure 3.2 – Length distribution of mature sequences in miRBase. We used miRBase v21 with a total of 35828 entries and we observed that the length is between 19-24 nt.

seq experiments. BrumiR takes as input sequencing files in FASTA or FASTQ formats. The sequencing data can be cleaned, using a fastq pre-processor [33] (*i.e.* fastp), to remove adapter sequences and trim low quality bases. BrumiR employs the Bcalm [35] tool to build a de Bruijn graph from the sRNA-seq reads. Bcalm uses a node-centric bi-directed de Bruijn graph where the nodes are k-mers, that is words of length k, and an arc between two nodes if the $k-1$ suffix of one node is equal to the $k-1$ prefix of the subsequent node, representing an exact overlap of $k-1$ bases [35]. A critical parameter of any de Bruijn graph approach is the k-mer size [50]. We observed that the length of all mature miRNA sequences stored in the miRBase database (v21) [97] have a minimum value of 18 nt (Figure 3.2). We thus empirically set the k-mer size equal to 18. Bcalm compacts the nodes of the de Bruijn graph into maximal unipaths by gluing all the nodes of the graph with an in-degree and an out-degree equal to one, thus generating the so-called unipath graph [35]. The unipath graph is the starting point of BrumiR (Figure 3.8.1). Notice that the unipath graph generated by Bcalm does not represent what is expected for a set of mature miRNAs (one connected component for each miRNA) and therefore further graph operations are needed. BrumiR uses a minimum k-mer frequency (KM value) of 5 and all k-mers with lower frequency are ignored, without losing most of the information contained in the sequencing reads (Figure 3.3).

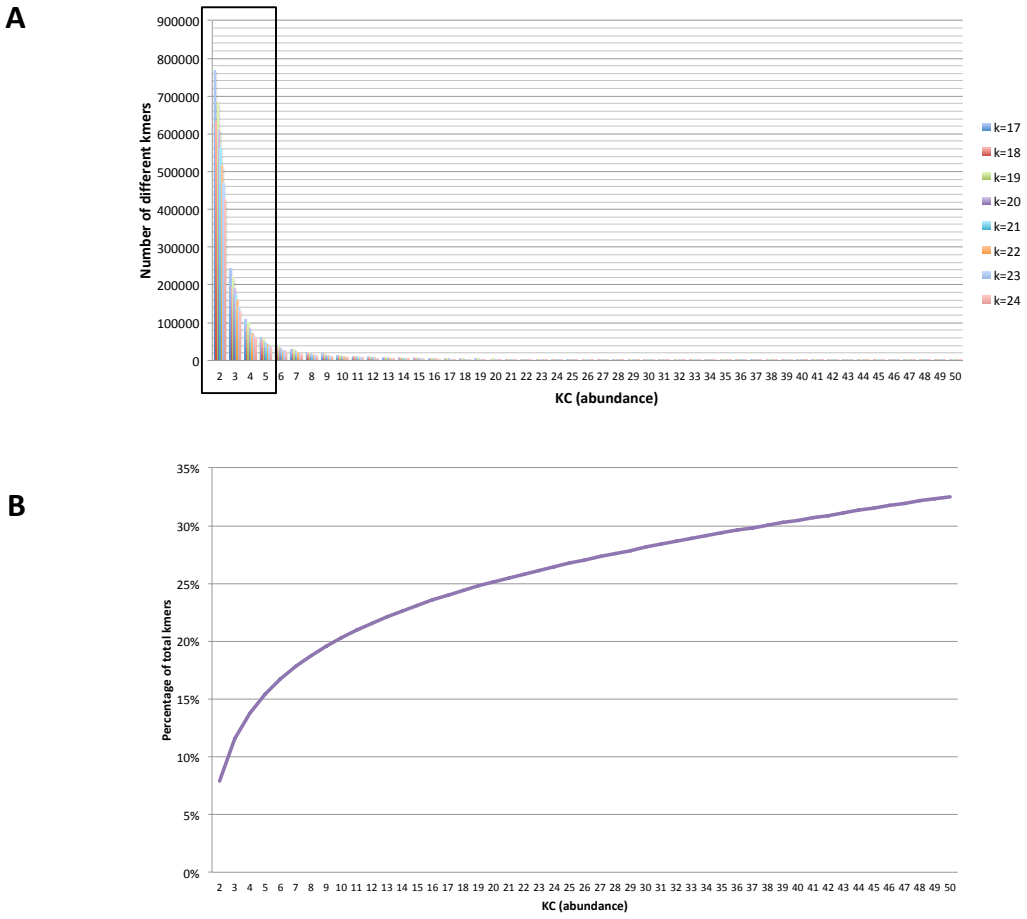

Figure 3.3 – Kmer spectrum of sRNA-seq data. A) The histogram shows the number of distinct kmers (Y-axis) as a function of the read coverage (KC X-axis). In the lower coverage of the spectrum (black rectangle), we observe a high number of distinct kmers which are likely sequencing errors. The kmers that correspond to noise represent approximately less than 15

3.3.2 Removing sequencing errors from the unipath sRNA-seq graph.

BrumiR deletes from the unipath graph all the nodes that have only one connection (degree equal to 1), known as dead-end paths or tips [36]. Usually, these nodes have a low abundance value associated to them (KM less than or equal to 5, the default parameter). Moreover, BrumiR deletes isolated nodes (degree equal to 0) having a low abundance; isolated nodes highly expressed are however conserved for further analysis. All these nodes are likely artifacts generated from sequencing errors because they are not deeply expressed in the sRNAs-seq reads [44]. BrumiR iterates this step 3 times in order to prune and clean the unipath graph (polishing). This operation, called 'tip removal', edits the original unipath graph, and therefore a new unipath graph with a new structure is generated (Figure 3.8.2).

3.3.3 An expressed mature miRNA has uniform coverage.

The unipath graph of a set of miRNAs from a sRNA-seq experiment has non-uniform coverage as different miRNAs and other elements may be connected in a single big component (Figure 3.8.1). BrumiR evaluates each connection of the unipath graph to identify those that link two nodes with a large expression difference. According to the miRNA biogenesis, after a stable miRNA precursor is cleaved by Dicer, among its three products, the miRNA mature sequence is the most abundant and when it is sequenced, it has a uniform expression along its sequence [61]. Thus due to miRNA biogenesis, it is possible to capture the complete miRNA mature sequence having a homogeneous expression [23] directly from the sRNA-seq experiments. BrumiR expects a similar KM value for k-mers originating from the same mature miRNA gene. Accordingly, if we observe two connected nodes that show a big difference in their abundance values, this connection is deleted and we keep the nodes unconnected. In particular, two unipaths $U = [a, b]$ connected in the graph have a KM value associated to them that represents their coverage from the reads information. BrumiR scans all the neighbor connections and if the difference between their KMs is larger than three-fold, the connection is deleted ($Ui_{km}/Uj_{km} > 3$). In this way, BrumiR defines a relative threshold that will depend on each unipath neighborhood in the graph. Finally, BrumiR repeats the tips removal step to eliminate new low frequency isolated nodes (Figure 3.8.4).

3.3.4 miRNAs and other sequences are captured in single connected components.

After the previous steps of BrumiR, a new unipath graph emerges, with a new structure. It is thus necessary to identify and classify the new connected elements within the graph (Figure 3.8). A connected component (CC) of a graph is a maximal strongly connected subgraph [112]. BrumiR computes the CCs of the unipath graph, and then each CC is

processed independently to identify miRNA candidates as well as to discard other sequences present in the unipath graph.

3.3.5 BrumiR classifies low abundance non-linear topologies as sequencing artefacts.

BrumiR detects topologies that are potentially related to sequencing errors and thus unlikely to be miRNA candidates. The shapes of these topologies were identified by visual inspection of several unipath graphs and are described in detail in Figure 3.4. Usually they have low KM and are composed of lowly expressed branching nodes with 3, 4 or 5 connections to the principal structures in the graph (Figure 3.4). Moreover, we observed that the sequences contained in these topologies were usually redundant and contained in other linear and more expressed CCs. In this way, we are not discarding relevant sequence information. BrumiR removes about 10% of the CCs in this step.

3.3.6 Re-assembling unipaths within each CC.

BrumiR re-assembles all unipaths present in the linear CCs by bundling the nodes with in and out degree equal to 1 into a new unipath. BrumiR classifies them into different types based on their length. The latter is the length of the sequence represented by the new unipath. All CCs having a length between 18 and 24 are stored as potential miRNA sequences. The CCs corresponding to an isolated node that have high KM (KM>50) are included in the latter group. CCs with lengths over 24 are classified as longer sequences or other types of genomic sequences captured along with the miRNAs. The longer sequences are put aside for later analysis. Moreover, BrumiR identifies circular CCs and branching CCs. The former are circular unipaths and the latter CCs with a high number of branching nodes. Branching CCs are not considered in the subsequent steps because they are likely sequencing errors (low abundance) or contamination present in the sRNA-seq data (Figure 3.5).

3.3.7 Re-clustering potential miRNAs.

After grouping unipaths by CCs, BrumiR builds an overlap graph to rescue the missing connections between potential miRNA candidates sharing an overlap with another candidate. First, BrumiR adds all the candidates as nodes of the overlap graph, then an all-vs-all k-mer comparison is performed using exact overlaps of length k=15. Candidates sharing an exact overlap are connected in the overlap graph. Then, the connected components are computed to identify clusters of miRNA candidates, and the most expressed candidate within each component is selected as the representative candidate of the cluster. The representative candidates are compared all-vs-all in a second overlap step that allows a maximum edit distance of 2, which is implemented using the edlib library [187]. BrumiR then builds a second overlap graph, computes again the connected components, and selects the most expressed candidate

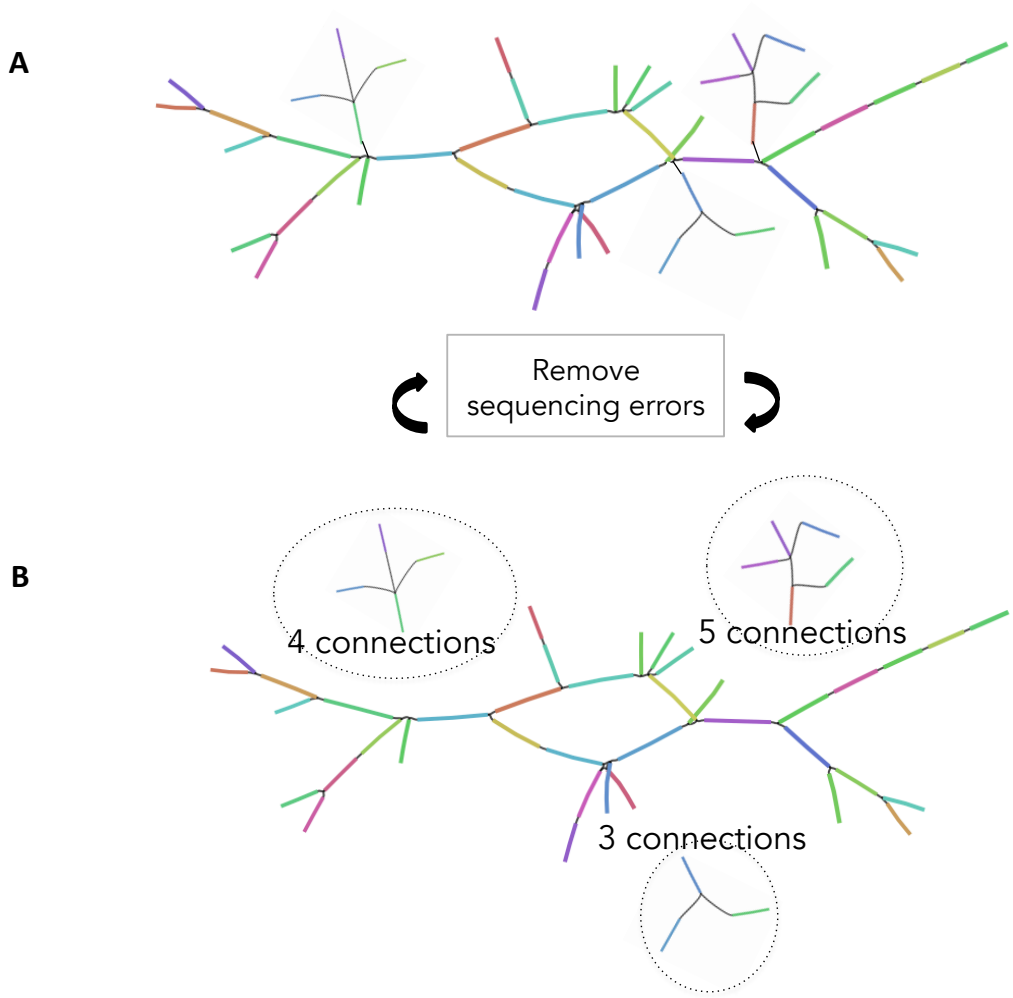

Figure 3.4 – BrumiR classifies low abundance non-linear topologies as sequencing errors. A) BrumiR identifies these topologies connected to the principal structures in the graph, which appear after the first tip removal steps of BrumiR. B) These topologies have low abundance (KM value) and are composed of branching nodes with 3, 4, or 5 connections.

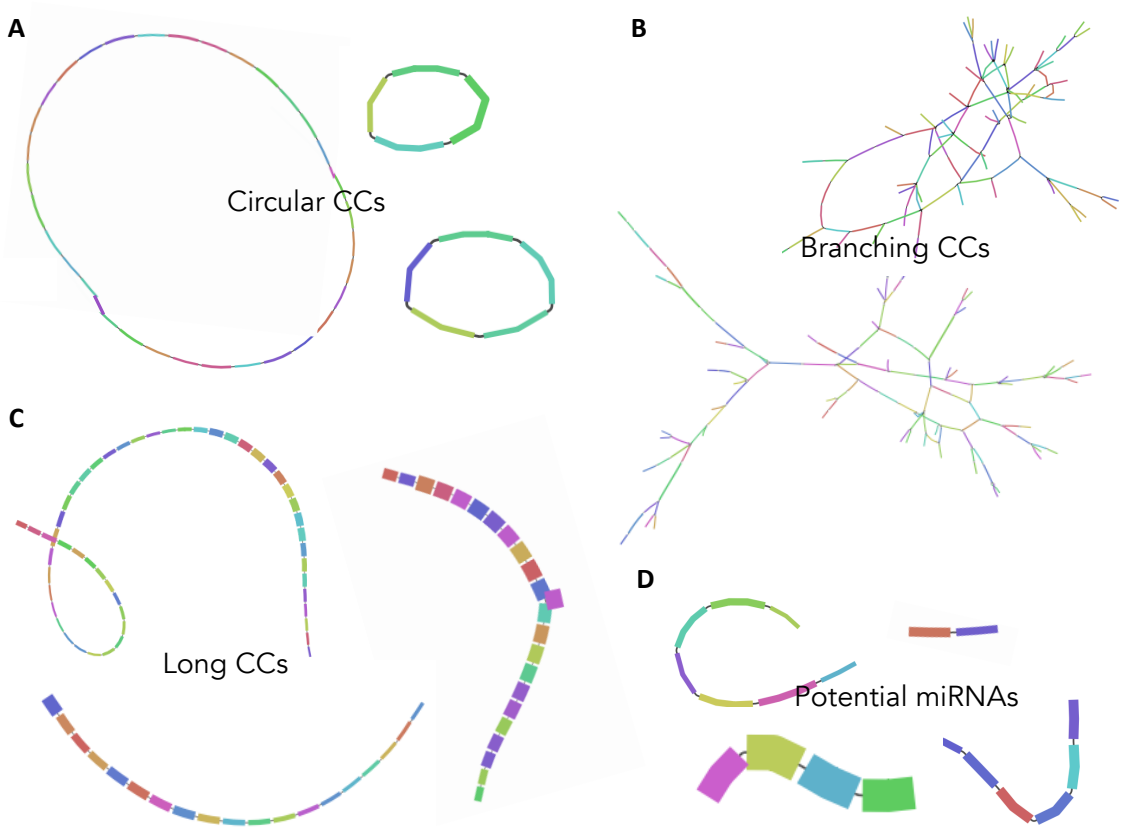

Figure 3.5 – Re-assembling unipaths within each CC. BRUMIR re-assembles all unipaths present in a linear CC by bundling the nodes with in and out-degree equal to 1 into a new unipath. BRUMIR rebuilds each unipath within a CC and classifies them into different types. A) Circular CCs: when all unipaths are have an in and out-connection, we classify the CC as a circular sequence that is not a putative miRNA. B) Branching CCs: when we detect a CC with a high number of branching nodes, we do not consider it anymore for the moment, because we consider it related to sequencing errors (usually they have a low KM value). C) Long CCs: when we detect more than 10 unipaths, we can classify them as longer non-coding sequences, but we still keep them for later analysis. D) Potential miRNAs: all assembled unipaths (CCs) having a length between 18 and 24 are stored as potential miRNA sequences.

as the representative of each cluster. The other members of each connected component are classified as putative isomiRs and saved in a file for later analysis.

3.3.8 Identifying other expressed RNA sequences.

In sRNA-seq experiments, different types of RNAs are expressed, some of which, such as small non-coding RNA elements, may have similar length with miRNAs [101]. The RFAM database [86] is a collection of curated RNA families including three functional classes of RNAs (non-coding, cis-regulatory elements, and self-splicing RNAs), which are classified into families according to their secondary structure and sequence information (Covariance Models) [86]. We downloaded 3,017 RNA families present in RFAM (v14.1) and excluded 529 miRNA families [87]. The sequences of 2,488 RFAM families were concatenated (a total of 2,736,549 sequences) and used to build a 16-mer database with the KMC3 k-mer counter tool [95] ("-fm -n100 -k16 -ci5"). All distinct 16-mers with a frequency lower than 5 were excluded, leading to a total of 6,204,556 distinct 16-mers related to RNA elements. Additionally, we downloaded all the mature miRNA sequences from miRBase (v22.1) [97] and built a 16-mer database with KMC3 ("-fm -n100 -k16 -ci1 mature.fa.gz"). RFAM 16-mers matching 16-mers from the 16-mer mature miRBase database were excluded from RFAM, leading to a 16-mer RFAM database with a total of 6,204,487 distinct 16-mers (Figure 3.6). Finally, the BrumiR candidates (18-24 length) were matched to the 16-mer RFAM database, and matching candidates were excluded and reported as sequences potentially associated to other RNA elements. The BrumiR candidates passing the aforementioned filter are reported as the final list of miRNA candidates.

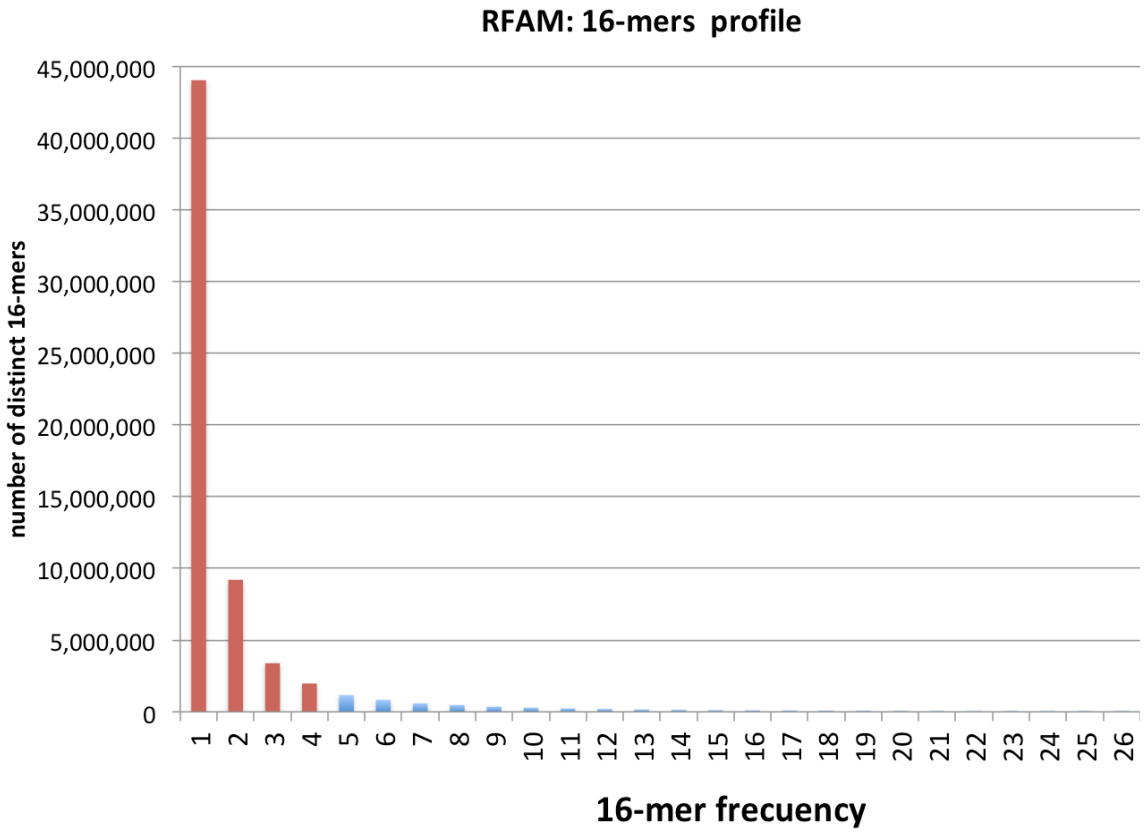

Figure 3.6 – 16-mer profile of RFAM entries.

3.4 BrumiR algorithm: from miRNA reads to a de Bruijn graph.

The main idea behind BrumiR is that mature miRNAs can be discovered directly from the information contained in the sequenced sRNA-seq reads. To achieve this, BrumiR starts by building a de Bruijn graph (using k-mers of size 18) from the sRNA-seq reads, then compacting all the simple nodes thus leading to the unipath graph [35] (Figure 3.8.1). The unipath graph encodes all the sequence information of the sRNA-seq experiment, including sequencing errors, adapters, and other types of sequences (Figure 3.8.1). The construction of the unipath graph allows to avoid entirely the alignment of the sRNA-seq reads to a reference genome. Following the unipath graph construction, BrumiR cleans the graph by removing tips (dead-end nodes) with low expression/abundance (KM < 5), which are usually generated from sequencing errors (Figure 3.8.2).

One feature of the miRNA biogenesis is that after Dicer cleavage, the mature miRNA is the most abundant of the three by-products and when it is sequenced, it has a uniform expression along its sequence [61]. Therefore, BrumiR expects that the neighbor elements within a particular putative miRNA will have similar expression. BrumiR checks all neighbor connections (arcs), and deletes any connection with a relative expression difference larger than 3 fold (Figure 3.8.3), and the new graph is cleaned again by removing tips (Figure 3.8.4). Clusters of unipaths (connected components) with topologies related to sequencing errors are also removed (Figure 3.8.5). BrumiR attempts to re-assemble all unipaths within a connected component (CC) of the graph, and those with between 18 and 24 nt are classified as putative miRNAs, while longer re-assembled unipaths (>24 nt) are classified as other longer sequences (Figure 3.8.6). BrumiR then restores missing connections by re-clustering the putative miRNAs performing an all-vs-all comparison. The most expressed miRNA is selected as the representative of the cluster (Figure 3.8.7) and the remaining members are classified as potential isomiRs (Figure 3.8.7).

The final BrumiR step uses the RFAM database to discard predicted miRNAs matching to other classes of RNA (*e.g.* Ribosomal genes, Figure 3.8.8). As an example, BrumiR reduces the input sRNA-seq data by five orders of magnitude generating less than 1,000 putative mature miRNAs (24 million input reads to 966 miRNA candidates, see Figure 3.8.10). Finally, BrumiR outputs several FASTA files with all predicted mature miRNAs, all longer RNAs, putative isomiRs, other sRNAs (RFAM comparison), and a table with expression values for each predicted miRNA (Figure 3.8.9). Additionally, BrumiR outputs the final graph in GFA format, which can be explored using Bandage [199] (Figure 3.8).

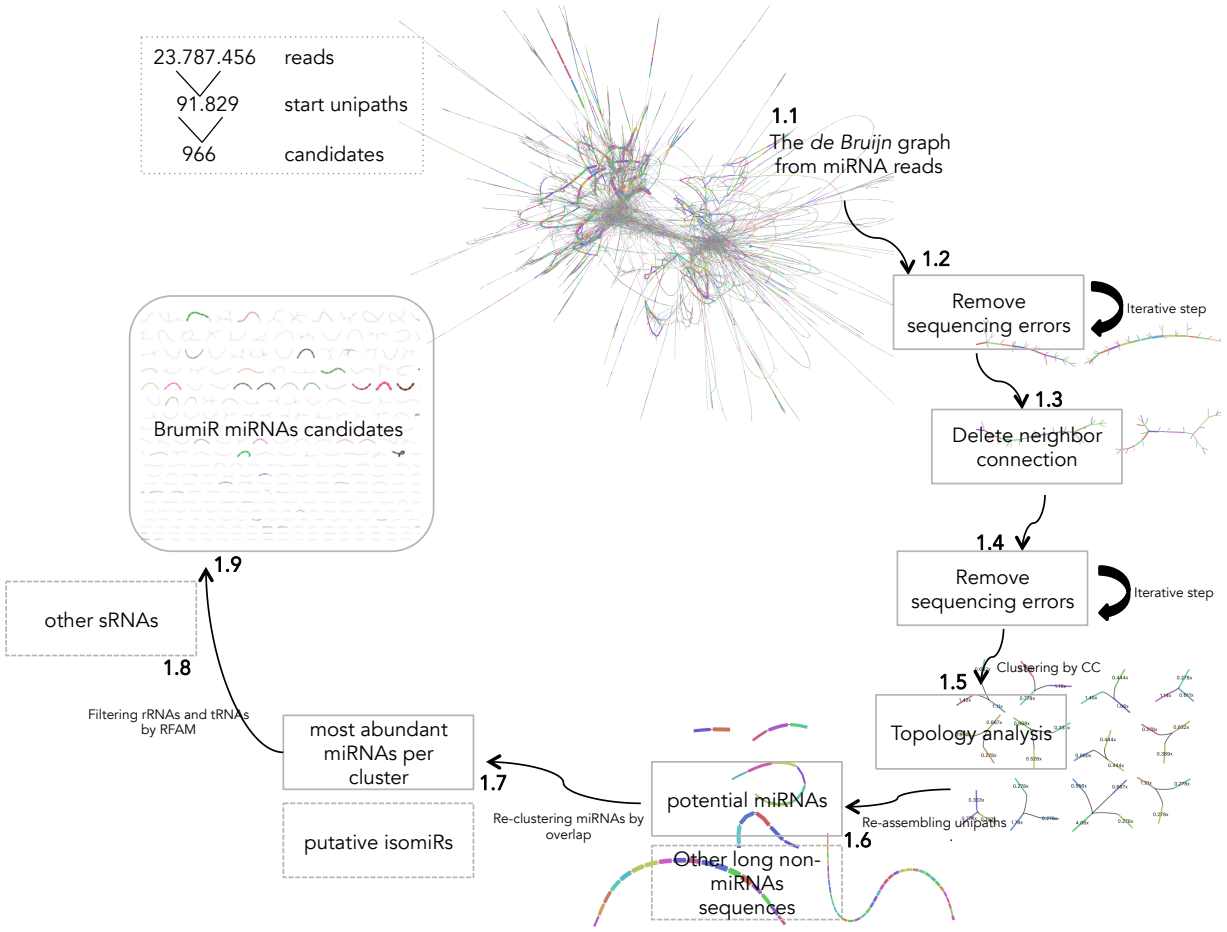

Figure 3.7 – **BrumiR algorithm**. Different steps of BrumiR to discover miRNAs from sRNA-seq data. 1.1 De Bruijn graph step, 1.2 Tips removal iterative step, 1.3 Delete neighbor connection step, 1.4 Tips removal step repetition, 1.5 Topology analysis step, 1.6 Re-assembling unipaths by CC step, 1.7 Re-clustering by overlap step, 1.8 Filtering other sRNAs by RFAM step, 1.9 BrumiR candidates catalog.

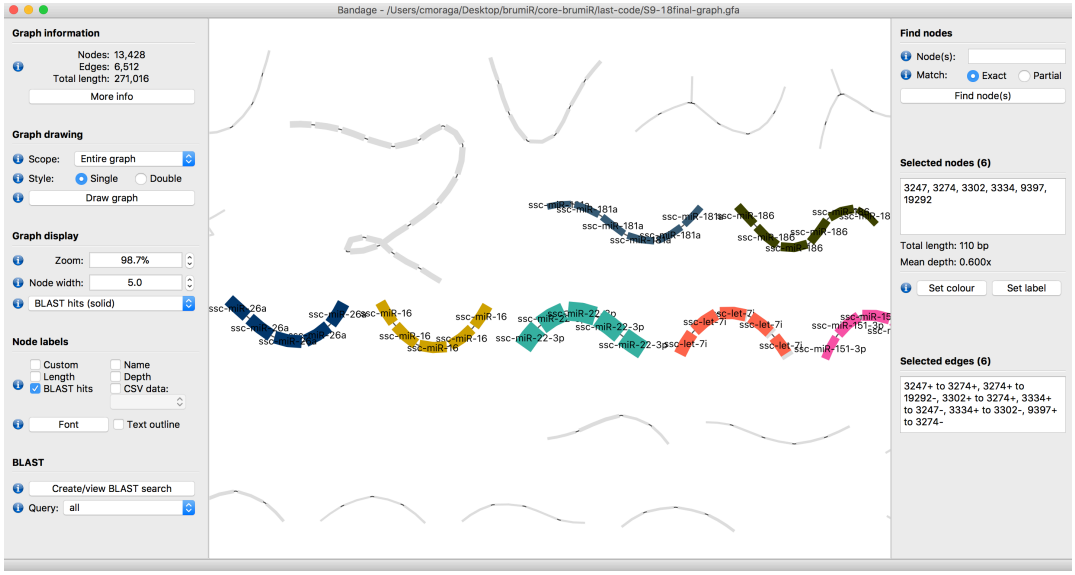

Figure 3.8 – Visualization with Bandage. BRUMIR provides an output compatible with the Bandage software, which can be employed to visualize and explore the results in a user-friendly way.

3.5 Results: BrumiR achieves the highest accuracy on simulated data.

We simulated synthetic reads from animal and plant species, and compared the results of BrumiR to those obtained with the miRDeep2 [63] and miR-PREFeR [110] tools. The sRNA-seq reads were simulated using miRsim (https://github.com/camoragaq/miRsim), a tool that we developed specifically for simulating sRNA-seq reads from a list of known miRNA mature sequences. miRsim is based on wgsim (https://github.com/lh3/wgsim), which is a widely used tool for simulating short Illumina genomic reads. miRsim includes functionalities specific of sRNA-seq reads such as variable depth/coverage and shorter read lengths. miRNA mature sequences were obtained from miRBase [97] for animal (High Confidence) and plant species. The animal species that we considered were: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Danio rerio*, and *Caenorhabditis elegans*, while the following plant species were included: *A. thaliana*, *Oryza sativa*, *Physcomitrella patens*, *Zea mays*, and *Solanum lycopersicum*. Table 3.1 provides further details (*i.e.* number of reads, number of mature miRNAs etc.) for each simulated dataset. miRDeep2 was run on the animal datasets with the default parameters providing the respective reference genome. Similarly, miR-PREFeR was run with the default parameters on the plant datasets. BrumiR was run with the default parameters on both the animal and plant datasets. The list of simulated miRNAs was considered as the ground truth, and precision, recall and F-Score quality metrics were computed to assess the performance of each discovery tool. The benchmark metrics were defined as follows:

$$recall = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F - Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}$$

where:
TP = true positive elements predicted as miRNAs present in the miRBase input list.
FP = false positive elements predicted as miRNAs but not present in the miRBase input list.
FN = false negative elements not predicted as miRNAs, but that were present in the miRBase input list.
To evaluate the performance of BrumiR, we applied it to discover mature miRNAs on simulated sRNA-seq reads from 10 animal and 10 plant species (Figure 3.9A). We compared BrumiR to the state-of-the-art genome-based miRNA discovery tools miRDeep2 [63] and miR-PREFeR [110], which were developed specifically for animal (miRDeep2) and plant (miR-PREFeR) species. For each tested species, we generated two synthetic datasets with different error-rates (0.01 and 0.02) using the miRsim tool implemented and provided by the BrumiR toolkit (https://github.com/camoragaq/miRsim), and the high-confidence miRNAs annotated in the miRBase database. A total of 20 datasets with an average of 11.5

million reads were simulated. The list of simulated miRNAs was considered as the ground truth, and benchmark metrics (Figure 3.9C) were computed to assess the performance of BrumiR and of the other software (Table 3.1).

BrumiR recovered more mature miRNAs than the others, on average 92% (as opposed to 41% and 64% for miRDeep2 and miR-PREFeR, respectively), and presented the highest average recall across all the simulated datasets (Figure 3.9B). BrumiR recovered more than 90% of the simulated mature miRNAs in 17 of the 20 simulated datasets (Figure 3.9B). In particular on the *H. sapiens*, *M. musculus* and *D. melanogaster* datasets, BrumiR recovered three times more candidates than miRDeep2 (Figure 3.9B). As concerns precision, BrumiR tended to generate more putative candidates than MiR-PREFeR (median 316 vs 264) and miRDeep2 (median 445 vs 308). The slightly higher number of BrumiR candidates resulted in lower average precision than miRDeep2 (0.59 vs 0.69) and MiR-PREFeR (0.63 vs 0.87). This is due to the fact that BrumiR does not use the hairpin structure filter employed by the other software. If we consider both precision and recall (F-Score), BrumiR was the top performer in 16 of the 20 datasets evaluated (Figure 3.9C). With animal species, BrumiR always reached a higher F-score than miRDeep2. With plant species, BrumiR was better or comparable to miR-PREFeR on most datasets, except for *Z. mays* and *P. pattens* where miR-PREFeR reached a higher F-Score (Figure 3.9C). In terms of computational time, BrumiR was the fastest method. In particular, BrumiR core was on average 30X faster than miRDeep2 and 10X times faster than MiR-PREFeR (Table 3.2). The speed of BrumiR relies on efficient alignment-free and graph-based approaches.

Overall, we showed with simulated data that BrumiR discovers putative mature miRNAs without a reference genome across different eukaryotic species achieving the highest accuracy and computational efficiency.

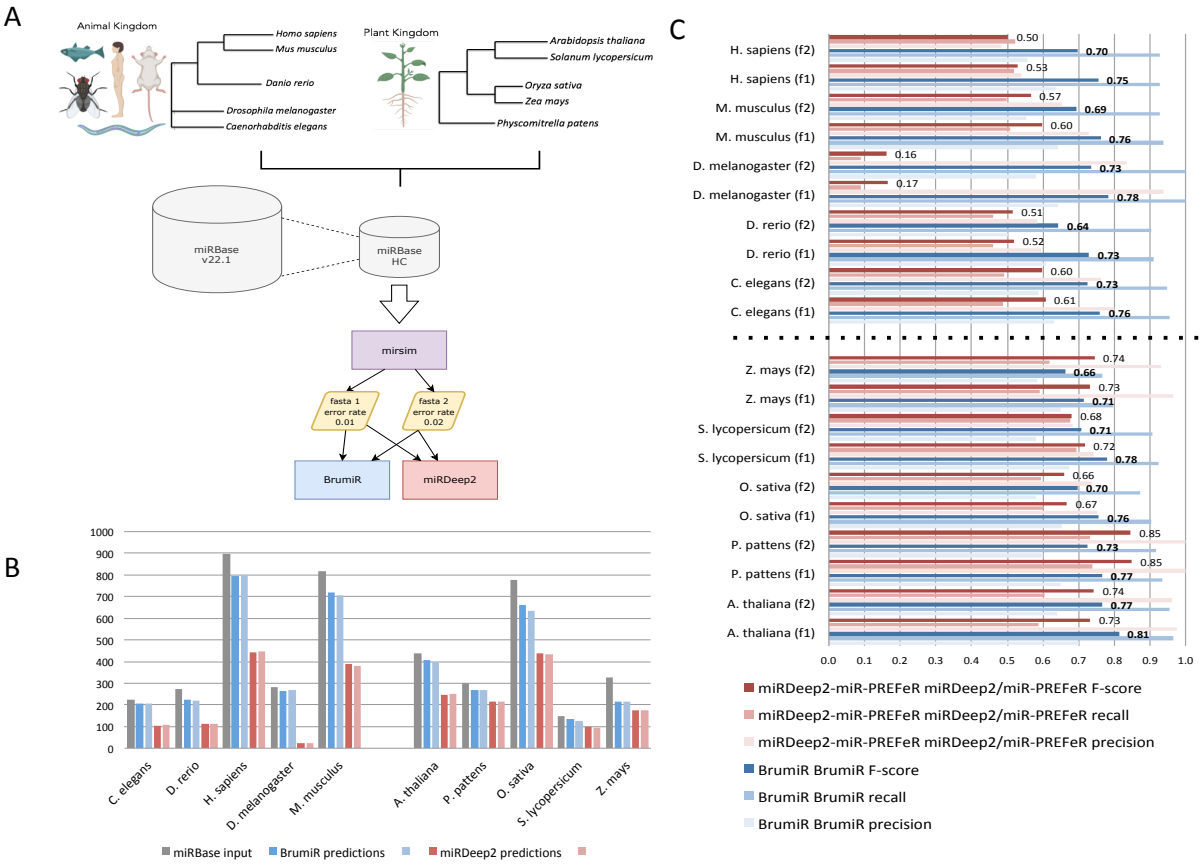

Figure 3.9 – **Synthetic benchmarking between BrumiR and miRDeep2**. A) Workflow and species selected, B) miRBase input vs miRNA true positive predictions for each tool (2 samples), C) Benchmarking metrics for all datasets tested.

Table 3.1 – Simulated sRNA-seq used to evaluate the performance of BrumiR.

						BrumiR			miRDeep2/miR-PREFeR			BrumiR			miRDeep2/miR-PREFeR		
		#reads (M)	mapped ref. genome	%mapped (ref. genome)	unipaths	candidates	TP	hit miRBase	candidates	TP	hit miRBase	precision	recall	F-score	precision	recall	F-score
Animal datasets																	
Homo sapiens	f1	22.91	21.01	91.71	79,986	1,205	766	794	1,136	614	443	0.64	0.93	0.75	0.54	0.52	0.53
	f2	22.91	20.46	89.32	108,829	1,391	776	798	1,298	627	449	0.56	0.93	0.70	0.48	0.52	0.50
Mus musculus	f1	21.29	19.90	93.49	73,003	1,117	716	719	700	509	389	0.64	0.94	0.76	0.73	0.51	0.60
	f2	21.29	19.39	91.07	99,995	1,284	710	707	767	501	380	0.55	0.93	0.69	0.65	0.50	0.57
Drosophila melagonaster	f1	7.46	7.13	95.56	26,785	418	269	265	32	30	24	0.64	1.00	0.78	0.94	0.09	0.17
	f2	7.46	6.90	92.43	36,683	472	274	270	36	30	24	0.58	1.00	0.73	0.83	0.09	0.16
Danio rerio	f1	7.00	6.37	91.02	24,073	337	205	224	306	182	113	0.61	0.91	0.73	0.59	0.46	0.52
	f2	7.00	6.16	88.05	33,370	396	198	221	309	180	113	0.50	0.90	0.64	0.58	0.46	0.51
Caenorhabditis elegans	f1	5.42	5.31	98.01	20,082	323	204	205	196	157	105	0.63	0.95	0.76	0.80	0.49	0.61
	f2	5.42	5.13	94.62	28,449	351	206	205	206	157	106	0.59	0.95	0.73	0.76	0.49	0.60
Plant datasets																	
Zea mays	f1	8.44	6.68	79.17	22,503	230	149	215	166	160	159	0.65	0.80	0.71	0.96	0.59	0.73
	f2	8.44	6.47	76.72	29,232	252	147	215	188	175	174	0.58	0.77	0.66	0.93	0.62	0.74
Physcomitrella patens	f1	7.84	6.57	83.82	22,247	299	194	270	184	184	213	0.65	0.93	0.77	1.00	0.74	0.85
	f2	7.84	6.33	80.76	29,796	333	200	267	184	184	213	0.60	0.92	0.73	1.00	0.73	0.85
Oryza sativa	f1	19.92	14.39	72.25	61,034	849	553	662	1,296	974	440	0.65	0.90	0.76	0.75	0.60	0.67
	f2	19.92	13.90	69.76	83,602	955	554	634	1,408	1,038	433	0.58	0.87	0.70	0.74	0.59	0.66
Solanum lycopersicum	f1	3.52	2.65	75.33	11,751	180	121	133	1,078	798	100	0.67	0.92	0.78	0.74	0.69	0.72
	f2	3.52	2.58	73.42	16,041	209	121	126	979	669	94	0.58	0.91	0.71	0.68	0.68	0.68
Arabidopsis thaliana	f1	11.22	9.33	83.15	32,464	510	358	405	249	243	246	0.70	0.96	0.81	0.98	0.59	0.73
	f2	11.22	9.01	80.37	43,174	571	365	399	278	267	253	0.64	0.95	0.77	0.96	0.61	0.74

Table 3.2 – Total elapsed time per tool (seconds) on synthetic datasets. The total elapsed time reported include only the core step of each algorithm.

	Species	dataset	BrumiR	miRDeep2/miR-PREFeR
animal	*Homo sapiens*	f1	111	6394
		f2	433	9661
	Mus musculus	f1	97	5101
		f2	383	8169
	Drosophila melagonaster	f1	25	922
		f2	61	1182
	Danio rerio	f1	22	1679
		f2	55	2067
	Caenorhabditis elegans	f1	16	917
		f2	38	1044
plant	*Zea mays*	f1	20	177
		f2	58	173
	Physcomitrella patens	f1	19	196
		f2	55	222
	Oryza sativa	f1	63	1661
		f2	227	1342
	Solanum lycopersicum	f1	12	1204
		f2	20	1193
	Arabidopsis thaliana	f1	31	363
		f2	106	362

3.6 Conclusion.

In the first part of this Chapter, in Section 3.4, I introduced the BrumiR core algorithm, preceded in Sections 3.2 and 3.3 by a description of each step used by BrumiR to navigate within the graph, remove errors and infer the miRNA candidates. There indeed, I explain in detail how I solve every challenge met by BrumiR using a de Bruijn graph to catch and organize the information contained in the sRNA-seq reads. Unlike other state-of-the-art tools, BrumiR does not rely on a reference genome, on the availability of close phylogenetic species, or on conserved sequence information. Instead, BrumiR starts from a de Bruijn graph encoding all the reads and is able to directly identify putative mature miRNAs on the generated graph. Along with discovering miRNAs, BrumiR assembles and identifies other types of small and long non-coding RNAs expressed within the sequencing data.
To benchmark BrumiR, in Section 3.5, I presented miRsim, a tool for simulating sRNA-seq reads from mature microRNA sequences (mirBase). miRsim is able to simulate miRNAs with SNPs and insertion/deletion (Indel) polymorphisms, as well as to simulate reads with uniform substitution sequencing errors. The benchmark shows that BrumiR achieves the highest accuracy on simulated data.
In summary, we observe a better performance of BrumiR as compared with the state-of-the-art tools, but it is necessary to further see the usefulness of BrumiR on real datasets and to compare its predictions with that of the other methods in such context. To that purpose, we present in the next chapter a validation of the performance of BrumiR by benchmarking it on real datasets. The code of BrumiR is freely available at https://github.com/camoragaq/BrumiR.

Chapter 4

Benchmarking and validating the predictions of BrumiR on a real dataset using BrumiR2reference

Contents

4.1 Introduction **87**

4.2 Implementation **88**

4.2.1 Benchmarking BrumiR using real sRNA-seq reads. 88

4.2.2 Identifying precursor sequences for the candidates of BrumiR (BrumiR2Reference) 89

4.2.3 miRNA discovery from *Arabidopsis* root samples. 90

4.3 Results **92**

4.3.1 The hairpin structure of mature miRNAs is found in most of the BrumiR candidates. 92

4.3.2 Discovering novel miRNAs from sRNA-seq data of *A. thaliana* roots using BrumiR. 94

4.4 Conclusion. **105**

4.1 Introduction

In this Chapter, we benchmarked BrumiR on animal and plant species using real datasets to see its performance in a real scenario. The benchmark results show that BrumiR is very sensitive, besides being the fastest tool, and its predictions were supported by the characteristic

hairpin structure of miRNAs. In order to improve the prediction of BrumiR and validate the miRNA candidates, we developed BrumiR2reference, a new mapping tool that performs an exhaustive search to identify and validate the precursor sequences when a reference genome is available. To establish the principal features of a hairpin miRNA sequence, we performed a complete analysis of all precursor sequences present in miRBase v21 [97] using the bpRNA tool [42]. In order to test the power of BrumiR in a known biological context, we applied BrumiR on a well-annotated organism to determine the potential to uncover putative novel miRNA candidates. We sequenced a total of 18 sRNA-seq libraries from different stages of root development of *Arabidopsis thaliana* and we used the BrumiR toolkit to analyze our data. We annotated three novel miRNAs involved in plant development, showing with a real experiment how BrumiR infers novel information even in highly annotated genomes.

4.2 Implementation

4.2.1 Benchmarking BrumiR using real sRNA-seq reads.

We downloaded publicly available sRNA-seq data for the plant and animal species listed in the synthetic benchmark, and two datasets for each species were included (Table 4.1). Additionally, we included mirnovo [195], a tool that can discover miRNAs without a reference genome. The predictions of BrumiR were benchmarked along with miRDeep2 (v2.0.1.2) [63] and mirnovo for the animal datasets. Similarly, miR-PREFeR [110] replaced miRDeep2 for the plant datasets (Table 4.1).

The stand-alone packages of BrumiR, miRDeep2 and miR-PREFeR were used to discover miRNAs in all datasets. The software mirnovo was run using its web version because the stand-alone package was not available and the developer recommends the use of the web version instead. The miRNA discovery was performed for each sample independently using default parameters for miRDeep2, miR-PREFeR and mirnovo. In particular, we used the scripts provided by miRDeep2 and miR-PREFeR to map the reads to the reference genome, and the predictions for these tools were performed on the resulting alignment files. The mirnovo predictions were done using the animal and plant universal panel respectively, as recommended when the reference genome is not available. BrumiR was run using the command line and parameters provided in commands section 4.3.2. Moreover, the predictions of BrumiR were refined using the BrumiR2reference tool on the available reference genome of the selected species (Table 4.1). Benchmark metrics (precision, recall, and F-Score) were computed as before but considering all the annotated mature sequences present in miRBase (v22.1) [97] as the ground-truth.

4.2.2 Identifying precursor sequences for the candidates of BrumiR (BrumiR2Reference)

As mentioned before, the biogenesis of miRNAs involves both nuclear and cytoplasmic processing. In the nuclear case, Drosha cleaves long primary transcripts releasing 60-150 nt pre-miRNAS (60-70 nt in animals, 60-150 nt in plants), which form a hairpin loop [128].

Nowadays, various methods exist which identify the miRNAs by mapping the sRNA-seq reads to a reference genome, thus identifying a precursor sequence, and then evaluate *in silico* the RNA secondary structure and its minimum free energy (MFE) [23].

Unlike such current state-of-the-art tools that perform miRNA discovery by mapping all the sRNA-seq reads to a reference genome, BrumiR first generates candidates by operating directly on the sRNA-seq reads. The reduced list of potential BrumiR miRNA candidates then permits the computation of a more exhaustive alignment than when mapping directly the sRNA-seq reads to the reference genome.

BrumiR aligns each candidate to the reference genome using an exact alignment method that computes the edit distance [145] between two strings and thus supports mismatches, insertions and deletions. The BrumiR2reference tool divides the reference genome in non-overlapping windows of 200bp (adjustable parameter), then the window is indexed using 12-mers and each miRNA candidate is matched in both strands (split at 12-mers). When a 12-mer match is found, an exhaustive alignment is computed between the window and the matching miRNA candidate. The alignment is performed using a fast implementation of Myers' bit-vector algorithm [187].

A miRNA candidate is stored as hit if the alignment in the current genomic window has an edit distance less than or equal to 2. After scanning all the genomic windows, the vector of hits is sorted by miRNA-candidate, edit distance (0-2), and alignment sequence coverage. For a single miRNA-candidate, a maximum of 100 genomic locations (best hits) are selected. BrumiR2reference then builds a potential precursor sequence for each selected hit using a strategy similar to the ones employed by miRDeep2 [63] and Mirinho [79]. BrumiR excises the potential precursor hairpin sequence from the flanking genomic coordinates of the reported miRNA candidate hits (mature sequence) in both strands. Potential precursor hairpin sequences of length 110 bp are built for animal species from both strands, while for plant species hairpin sequences of lengths 110, 150, 200, 250 and 300 bp are built from both strands [145]. Secondary structure prediction for all the potential precursor sequences is performed using RNAfold (v2.4.9) [125]. Secondary structures with a minimum free energy in the range of 15-80 kcal/mol are checked for a hairpin loop characteristic of miRNAs [172] (Figure Figure 4.1). Structures with a hairpin loop composed of a single segment without pseudo-knot, multi-loops, external loops and with less than 5 bulges, 3 dangling ends, and 10 internal loops are classified as characteristic secondary structures of miRNA precursor sequences. The aforementioned filters were derived from analyzing the secondary structure of 38,589 precursor sequences stored in miRBase (v22.1) [97] using a modified version of the

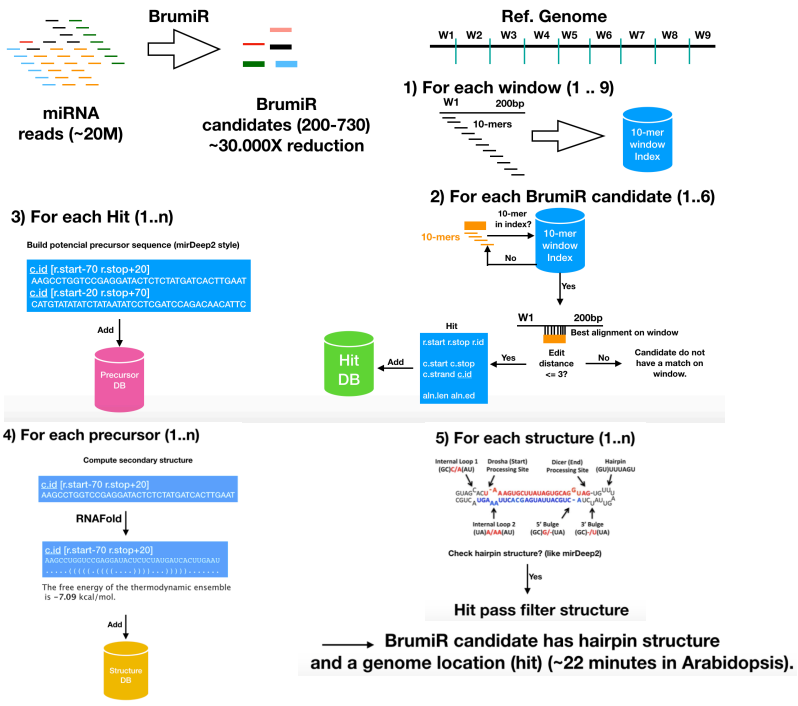

Figure 4.1 – Workflow of the BrumiR2Reference tool. The main steps involved the mapping of the miRNA candidates to the genome using non-overlaping windows (1); each hit is further refined using an exhaustive alignment (2). For each hit, a precursor sequence is built (3), and its secondary structure is determined using RNAfold (4). Finally, structures fulfilling a set of criteria (5) are classified as precursor sequences.

bpRNA program [42] (Figure 4.2).

4.2.3 miRNA discovery from *Arabidopsis* root samples.

A. thaliana Col-0 seedlings were grown hydroponically on Phytatrays on 0.5X Murashige and Skoog medium (Phytotechnology Laboratories, cat. M519) under long-day conditions (16h light and 8h dark) at 22°C. Total RNA was isolated from plant roots after 5, 9, 13, 17, 21, and 25 days post-germination using the mirVana miRNA Isolation Kit (Thermo Fisher Scientific, cat. AM1560). RNA concentration was determined using the Qubit RNA BR Assay Kit (Thermo Fisher Scientific, cat. Q10210), and integrity was verified by capillary electrophoresis on a Fragment Analyzer^{TM} (Advanced Analytical Technologies, Inc.). The indexed sRNA libraries were built employing the TruSeq small RNA Sample Preparation Kit (Illumina, Inc.) following the manufacturer's instructions. Briefly, 3' and 5' adaptors were sequentially ligated to 1 μg of total RNA prior to reverse transcription and library amplification by PCR. Size selection of the sRNA libraries was performed on 6% Novex TBE PAGE Gels (Thermo Fisher Scientific, cat. EC6265BOX) and purified by ethanol precipitation. Both the library size assessment and library quantification were carried out in a Fragment Analyzer^{TM}. Finally,

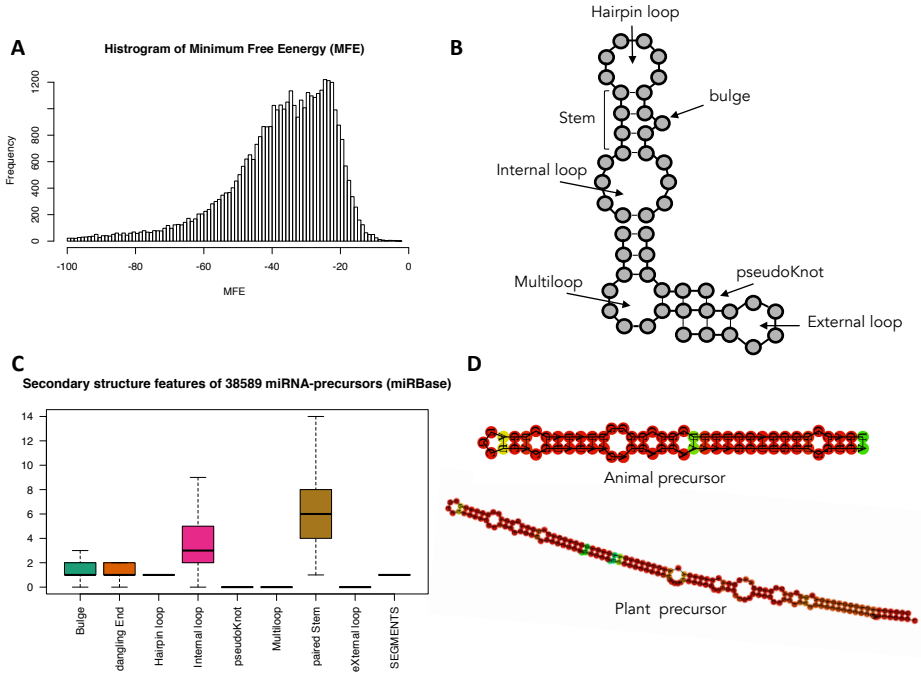

Figure 4.2 – Structure properties of miRBase precursor sequences. A) Free-energy distribution of 38.589 precursor sequences folded with RNAFOLD. B) Different types of RNA secondary structure elements composing precursor miRNA sequences. C) Analysis of secondary structure elements performed on 38.589 precursor sequences in miRBase using the BPRNA package. D) Examples of precursor sequences for animal and plant species.

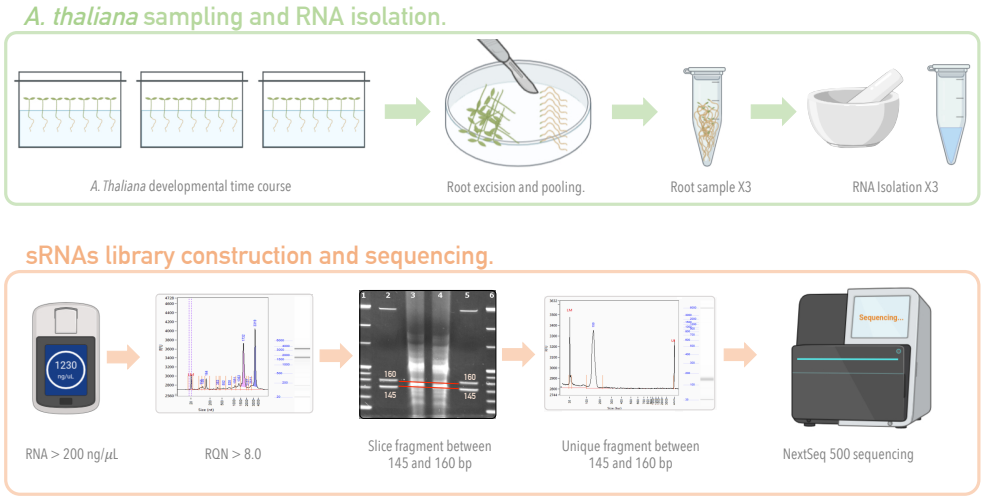

Figure 4.3 – Applying BrumiR on SRNA-seq from *Arabidopsis* root libraries. A) Experimental procedure to obtain the samples from the root. B) Construction of the sRNA-seq libraries to be sequenced later.

the libraries were pooled and sequenced on an Illumina NextSeq 500 platform (Figure 4.3). All samples were analyzed with BrumiR separately with default parameters to identify the candidate miRNAs. We further validated the candidates having a putative precursor with a hairpin structure analysis using the BrumiR2reference tool with the reference genome for *A. thaliana* (GCF_000001735.4_TAIR10.1_genomic.fna). All validated candidate miRNAs were compared with known miRNAs described for *A. thaliana* (437) present in miRBase (v21). The putative novel miRNAs were curated manually. Specifically, we checked the hairpin features, mature sequence alignment position, star sequence in the precursor sequence, mismatches in the seed region, and exact overlap in the antisense miRNA sequence [23]. Then a target analysis was performed using the Araport 11 cDNA library with the plant-specific psRNATarget algorithm (based on a best expectation score) [41].

4.3 Results

4.3.1 The hairpin structure of mature miRNAs is found in most of the BrumiR candidates.

In order to assess the performance of BrumiR on real data, we collected public datasets for the same plant and animal species evaluated in the synthetic benchmark (Figure 3.9A). On average, 15.4 and 18.2 raw million reads were used for the animal and plant datasets (Table 4.1), respectively. The predictions of BrumiR were compared against those of the state-of-the-art tools encompassing reference and *de novo* based methods [195, 63, 110]. In

particular, we included MIRNOVO that similarly to BRUMIR can discover mature miRNAs directly from the reads. Before running the tools, low-quality reads were removed using FASTP [33] (∼10%). All the predicted miRNAs for each tool were annotated using the miRBase database to identify known and novel predictions. On average, BRUMIR predicted ∼1,000 putative mature miRNAs for the animal species, which was ∼2.7X higher than the MIRDEEP2 candidates and 1.7X lower than the candidates predicted by MIRNOVO (Figure 4.4A1). For plant species, BRUMIR predicted on average ∼1,900 putative mature miRNAs, which was lower than the candidates predicted by MIR-PREFER [160], and higher than the predictions of MIRNOVO (301 on average) (Figure 4.4A1). A comparison using the miRBase [97] annotated miRNAs revealed that BRUMIR shared more candidates with MIRDEEP2 and MIR-PREFER than with MIRNOVO (Figure 4.4A2). However, an important fraction (on average more than 70%) of the miRBase-annotated candidates were exclusive to each tool (Figure 4.4A2), which summarizes the complexity of miRNA discovery.

Considering miRBase-annotated candidates as the ground-truth, we computed precision, recall and F-Score for all the evaluated tools (Figure 4.4B. BRUMIR achieved an accuracy (F-Score) better (animals) or comparable (plants) to the one obtained by the other software (Figure 4.4B3). Moreover, BRUMIR consistently reached the highest recall for most of the datasets evaluated (Figure 4.4B2). The precision values of BRUMIR were slightly lower for some datasets (Figure 4.4B1). However, none of the methods performed well on this metric (Figure 4.4B3). In particular, MIRDEEP2 reached the highest precision (∼0.7) on animal species and all methods performed poorly on the evaluated plant species (average precision <0.3). The low precision with plant species may be the product of a low number of entries annotated in miRBase for plants (10.414 vs 38.471 animals) as well as of a higher complexity of plant miRNAs [145]. The BRUMIR toolkit also provides a tool to determine the hairpin loop of miRNA precursor sequences, which is the main structural feature of miRNAs [172]. BRUMIR2REFERENCE maps the mature miRNA predicted by BRUMIR to the reference genome using an exhaustive alignment, generates precursor sequences, computes their secondary structure, and checks the hairpin structure using a variety of criteria inferred from analyzing more than 30,000 miRBase precursor sequences from animal and plant species. We used BRUMIR2REFERENCE as a double validation for all the predicted mature miRNAs generated by BRUMIR for the animal and plant datasets (Figure 4.4C). On average, BRUMIR2REFERENCE identified a valid precursor sequence having the characteristic hairpin structure for over 70% of the BRUMIR candidates (Figure 4.4C). In terms of speed, BRUMIR core was the fastest tool. BRUMIR was on average 120X and 220X times faster than MIRDEEP2 and MIR-PREFER, respectively (Table 4.2).

Overall, we demonstrated that BRUMIR is a competitive tool for discovering mature miRNAs without a reference genome. We showed that it was the most sensitive on most of the datasets tested. The performance of our method was not only faster, but also better or comparable to the state-of-the-art tools. Moreover, we also provided a new mapper approach to be used when a reference genome is available, to further verify if a precursor sequence of the predicted

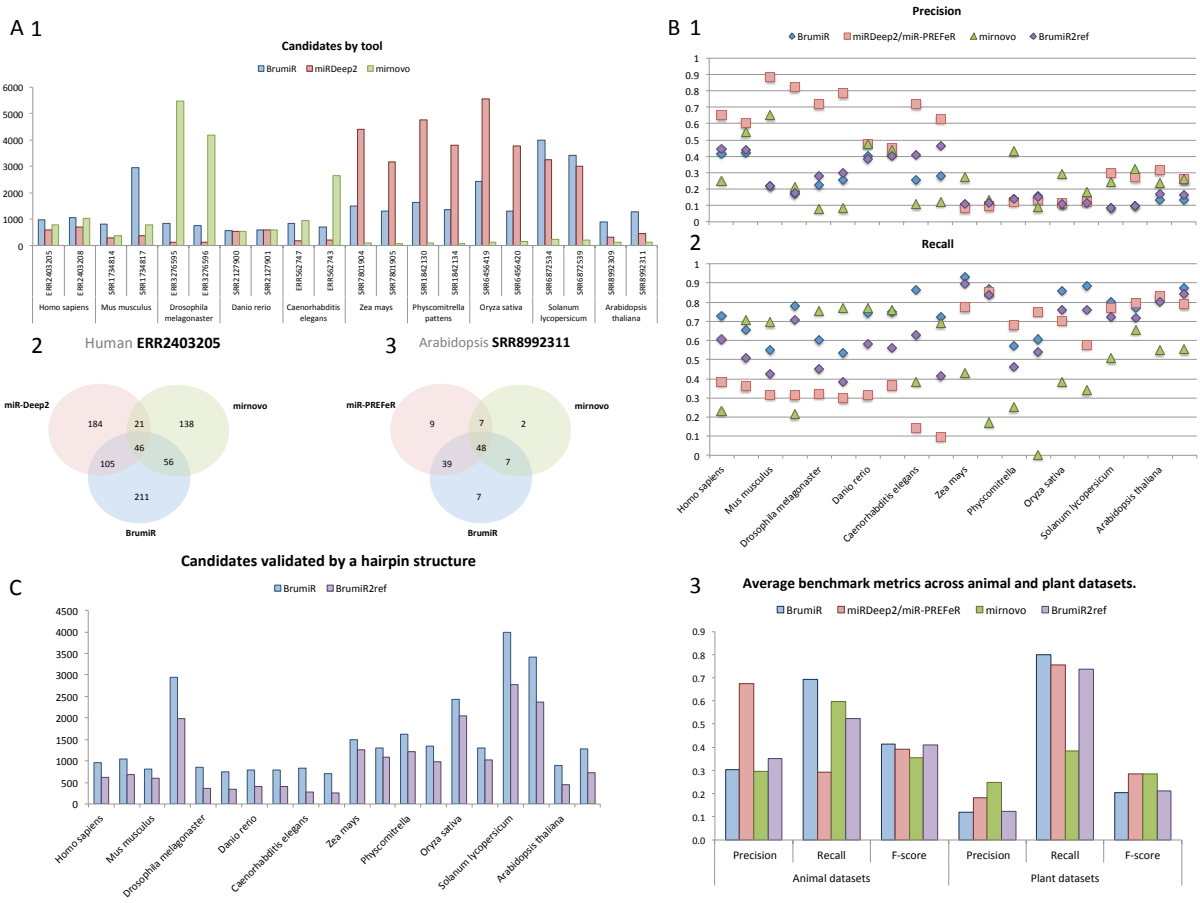

Figure 4.4 – **Real dataset benchmark of BrumiR and state-of-the-art tools**. A) Number of predictions by tool for all the datasets and the overlap between them for 2 datasets (1 for animal and 1 for plant); B) Benchmarking metrics computed using miRBase annotated miRNAs, precision and recall for each dataset; and average metrics, including F-score. C) BrumiR candidates validated by Hairpin structure (BrumiR2Reference).

mature miRNA is present in the genome. BrumiR therefore represents a reliable alternative for the discovery of mature miRNAs in model and non-model species with or without a reference genome.

4.3.2 Discovering novel miRNAs from sRNA-seq data of *A. thaliana* roots using BrumiR.

A. thaliana is one of the best characterized model organisms, and the first plant species in which miRNAs were cloned and sequenced [168]. To date, 436 mature miRNA sequences are included in the miRBase database. Most of these miRNAs have been identified by studies addressing the sRNAome of different plant organs [53], cell types [25], or responses to biotic

or abiotic stress using sRNA-seq [82, 150].

We sequenced sRNA-seq libraries from the roots of *A. thaliana* after different time points during vegetative development (Figure 4.3) to demonstrate the potential of BrumiR to discover novel mature miRNAs in a known biological context. BrumiR was run independently for each condition and replicate. The day 5 samples were excluded because of the low number of reads when compared to the other samples (Table 4.3). BrumiR predicted, on average, 1,120 mature miRNAs per sample, which were further refined to 678 using the BrumiR2reference tool. To take advantage of our experimental design, we considered as a putative miRNA the ones present in the three replicates (core predictions)[13] (Figure 4.5A). Novel miRNAs were identified using the following steps. First, predictions were classified as known miRNAs by comparing with miRBase (160 known miRNAs out of a total of 436 miRNAs already described for *A. thaliana* in miRBase). These known miRNAs were put aside to explore the sensitivity of BrumiR in detecting novel putative miRNAs. We then clustered the remaining putative miRNAs into three stages: early, late, and constitutive (Figure 4.5B). The days 9, 13 and 17 represent an early stage of the plant development [175]; days 17, 21 and 25 represent a late stage of the plant development [175], and the putative miRNAs expressed in all conditions represent the constitutive category. A total of 25 putative novel miRNAs were identified, and a manual curation was carried out using all the information provided by BrumiR.

Three curated novel miRNAs fulfilling all the recommended criteria to annotate miRNAs in plants [13] were discovered by BrumiR (Table 4.4). One of the curated miRNAs is located in chr1:29,612,248-29,612,361 (from now on denoted as miR-8) with a free energy of -40.5 and the characteristic hairpin structure of plant miRNAs (Figure 4.5C). Interestingly, this miRNA locus has not been previously discovered because its mature sequence maps to multiple chromosomes, and is therefore discarded by genome-based tools [221].

In an exploratory analysis to shed light on the potential targets of these novel miRNAs, we conducted an *in silico* target transcript prediction using the psRNATarget algorithm [41] (Table 4.5). FSD-1 (AT4G25100) was found to be one of the top genes regulated by this novel miRNA miR-8. In *A. thaliana*, FSD-1 encodes a Fe superoxide dismutase enzyme which regulates reactive oxygen species (ROS) levels of chloroplast and cytosol and participates in salt stress tolerance [73]. Moreover, knockout mutants of FSD-1 exibit a lower number of lateral roots, thereby suggesting an important role in root development [99]. FSD-1 is developmentally regulated, abundantly expressed from the 3rd day to the 13th day but significantly decreased in the following days, and its differential accumulation between root zones is related to emerging patterns of lateral roots and hair formation from trichomes [52]. Another predicted mRNA target of miR-8 is PER24 (AT2G39040), which is a peroxidase gene involved in the detoxification of ROS in the extracellular and Na+ homeostasis and which plays an important role in the resistance to salinity stress as does the FSD-1 gene [73]. It is plausible to say that these novel miRNAs may be involved in the fine-tuning of lateral root growth in the early stages of development. These results highlight the value of the BrumiR toolkit for discovering novel miRNA candidates with functional impact on the organisms studied, even

in the case where high quality genomes are available.

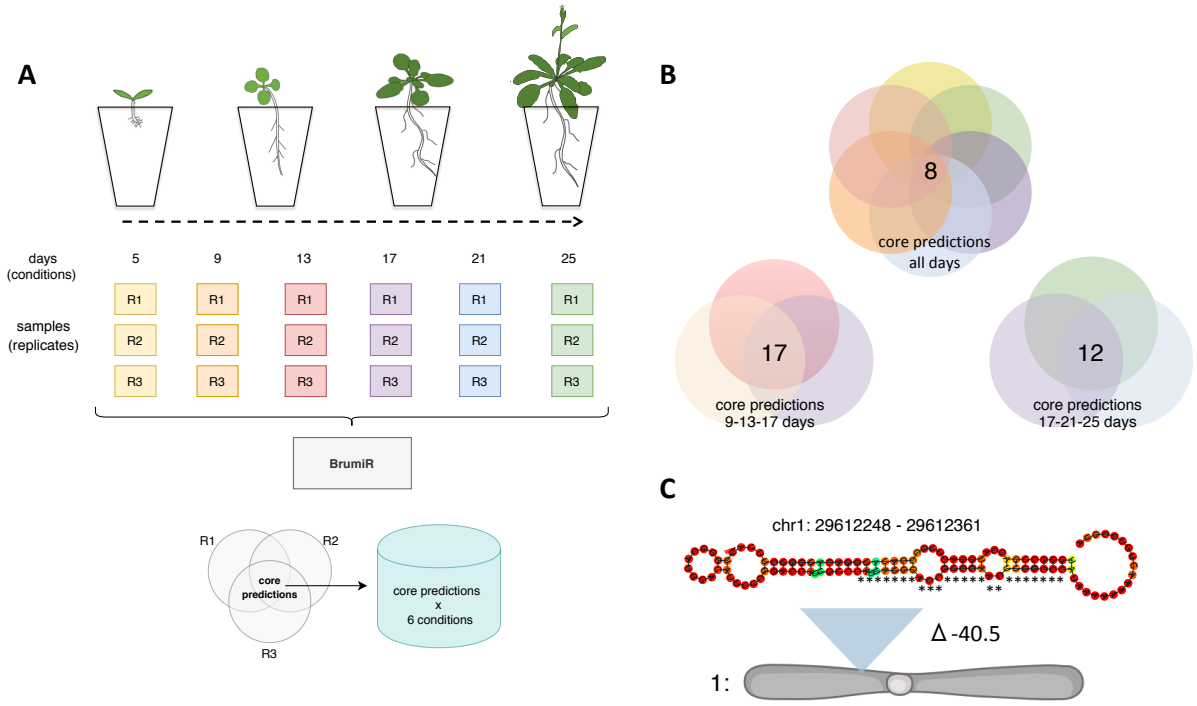

Figure 4.5 – **Applying BrumiR on sRNA-seq from *Arabidopsis* root libraries**. A) Experimental design implemented; roots from *Arabidopsis* in a time-scale per day as conditions were sequenced in three technical replicates. BrumiR was used to analyze all sRNA-seq librairies, and conserved predictions by the three replicates was considered as a core by condition. B) Different combinations of root growth per day were analyzed together to identify novel putative miRNAs conserved in all conditions. C) miR-8 is discovered as a novel miRNA, supported by the hairpin analysis, and is conserved in all replicates in all conditions.

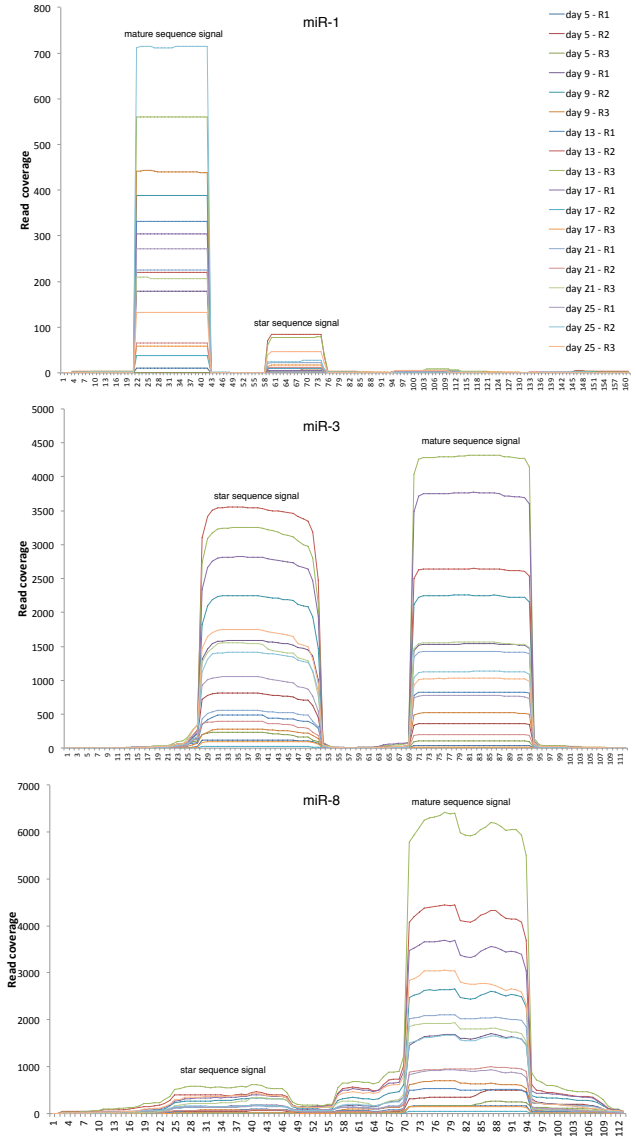

Figure 4.6 – Read coverage of novel precursor candidates found in the roots of *Arabidopsis thaliana*. Reads of each experimental condition were mapped back to the novel precursor sequences using Bowtie. Samtools was used to compute the read coverage (depth) along each precursor sequence. The three novel precursor sequences show the read signature peaks.

Table 4.1 – Real sRNA-seq data used to evaluate the performance of BrumiR.

		Homo sapiens		*Mus* musculus		*Drosophila* melagonaster		*Danio* rerio		*Caenorhabditis elegans*		*Zea mays*		*Physcomitrella patens*		*Oryza sativa*		Solanum *lycopersicum*		Arabidopsis *thaliana*	
		ERR2403205	*ERR2403208*	*SRR1734814*	*SRR1734817*	*ERR3276595*	*ERR3276596*	*SRR2127900*	*SRR2127901*	*ERR562747*	*ERR562743*	*SRR7801904*	*SRR7801905*	*SRR1842130*	*SRR1842134*	*SRR6456419*	*SRR6456420*	*SRR6872534*	*SRR6872539*	*SRR8992309*	*SRR8992311*
	# reads (M)	23.79	23.20	3.49	39.44	15.42	11.77	5.96	5.81	18.01	7.11	14.96	15.96	8.26	7.38	14.10	13.19	21.09	18.21	36.78	32.83
	T. reads (M)	21.84	20.25	3.48	39.35	15.30	11.69	5.77	5.53	10.72	6.67	11.70	11.77	7.97	7.14	13.26	11.58	19.83	17.34	36.07	32.51
	%Ref. mapped	70.51	76.39	48.41	48.17	51.83	56.56	69.80	64.79	82.62	86.91	23.36	29.29	51.39	44.71	67.48	56.65	56.20	58.60	93.75	91.83
	Unipath (k)	91.8	94.3	63.3	479.5	73.1	53.9	43.7	48.8	43.5	77.3	174.9	168.0	66;9	61.3	290.1	187.2	506.7	487.2	151.2	148.4
BrumiR	candidates	966	1046	813	2954	847	743	569	579	824	696	1501	1299	1628	1350	2435	1301	3992	3405	899	1282
	TP	401	438	175	497	190	190	230	237	212	192	156	149	228	212	265	164	331	319	123	169
miRDeep2	candidates	579	711	278	383	120	115	537	579	171	196	4401	3158	4769	3798	5569	3774	3257	3001	311	441
/ miR-PREFeR	TP	378	428	245	315	86	90	255	264	123	124	376	301	588	507	637	456	975	812	98	115
mirnovo	candidates	789	1017	358	785	5476	4181	537	597	955	2642	91	82	90	78	127	147	230	210	120	134
	TP	196	556	234	166	439	365	255	260	100	331	25	11	39	7	37	27	56	67	29	35
	precision	0.42	0.42	0.22	0.17	0.22	0.26	0.40	0.41	0.26	0.28	0.10	0.11	0.14	0.16	0.11	0.13	0.08	0.09	0.14	0.13
BrumiR	recall	0.73	0.66	0.55	0.78	0.60	0.54	0.74	0.75	0.86	0.72	0.93	0.87	0.57	0.61	0.86	0.88	0.80	0.77	0.83	0.87
	F-score	0.53	0.51	0.31	0.28	0.33	0.35	0.52	0.53	0.40	0.40	0.19	0.20	0.22	0.25	0.19	0.22	0.15	0.17	0.24	0.23
miRDeep2	precision	0.65	0.60	0.88	0.82	0.72	0.78	0.47	0.46	0.72	0.63	0.09	0.10	0.12	0.13	0.11	0.12	0.30	0.27	0.32	0.26
/	recall	0.38	0.36	0.32	0.31	0.32	0.30	0.31	0.36	0.14	0.10	0.78	0.86	0.68	0.75	0.70	0.58	0.77	0.80	0.83	0.79
miR-PREFeR	F-score	0.48	0.45	0.47	0.45	0.45	0.43	0.38	0.40	0.24	0.17	0.15	0.17	0.21	0.23	0.20	0.20	0.43	0.40	0.46	0.39
	precision	0.25	0.55	0.65	0.21	0.08	0.09	0.47	0.44	0.10	0.13	0.27	0.13	0.43	0.09	0.29	0.18	0.24	0.32	0.24	0.26
mirnovo	recall	0.23	0.71	0.69	0.22	0.75	0.77	0.77	0.76	0.38	0.69	0.43	0.17	0.25	0.00	0.38	0.34	0.50	0.66	0.55	0.55
	F-score	0.24	0.62	0.67	0.21	0.14	0.16	0.59	0.55	0.16	0.21	0.34	0.15	0.32	0.01	0.33	0.24	0.33	0.43	0.34	0.35
	precision	0.44	0.44	0.21	0.18	0.28	0.30	0.38	0.40	0.41	0.46	0.11	0.11	0.14	0.15	0.11	0.12	0.09	0.09	0.17	0.16
BrumiR2ref	recall	0.60	0.51	0.42	0.71	0.45	0.38	0.58	0.56	0.63	0.41	0.90	0.84	0.47	0.54	0.76	0.76	0.72	0.72	0.81	0.84
	F-score	0.51	0.47	0.29	0.28	0.35	0.34	0.46	0.47	0.50	0.44	0.19	0.20	0.21	0.24	0.19	0.20	0.16	0.17	0.28	0.27

Table 4.2 – Total elapsed time per tool (seconds) on real datasets. The total elapsed time reported include only the core step of each algorithm.

	Species	SRA_ID	BrumiR	miRDeep2/miR-PREFeR
animal	*Homo sapiens*	ERR2403205	48	6319
		ERR2403208	50	5421
	Mus musculus	SRR1734814	23	8342
		SRR1734817	333	11103
	Drosophila melagonaster	ERR3276595	31	5382
		ERR3276596	27	8270
	Danio rerio	SRR2127900	20	4921
		SRR2127901	22	5554
	Caenorhabditis elegans	ERR562747	24	10159
		ERR562743	28	8172
plant	*Zea mays*	SRR7801904	59	52875
		SRR7801905	50	13002
	Physcomitrella patens	SRR1842130	37	25909
		SRR1842134	30	39644
	Oryza sativa	SRR6456419	124	34346
		SRR6456420	66	17916
	Solanum lycopersicum	SRR6872534	378	31902
		SRR6872539	334	37629
	Arabidopsis thaliana	SRR8992309	50	4045
		SRR8992311	49	1802

Table 4.3 – miRNA discovery from the root samples of *Arabidopsis thaliana* using BrumiR.

		Raw reads (M)	Processed reads (M)	Unipaths	Candidates	Hairpin validated	Core predictions	Known miRNAs	Putative novel miRNAs
	1	22.19	2.10	16,122	117	72			
day 5	2	25.31	1.35	7,077	48	34	45	36	5
	3	25.77	2.97	18,687	116	69			
	4	24.65	15.60	90,588	989	592			
day 9	5	21.70	14.58	61,374	685	394	295	78	141
	6	24.66	14.46	107,404	1,698	1,045			
	7	29.30	19.63	94,222	1,153	583			
day 13	8	25.73	14.49	96,656	1,238	757	468	88	258
	9	27.94	20.22	144,949	1,944	1,221			
	10	21.77	13.44	61,014	738	441			
day 17	11	19.78	10.33	36,673	450	245	238	86	96
	12	16.49	9.55	89,471	1,264	790			
	13	24.42	17.80	127,221	1,852	1,135			
day 21	14	16.81	6.90	34,903	480	312	212	89	80
	15	22.54	5.54	26,271	434	267			
	16	24.19	16.31	133,004	2,069	1,273			
day 25	17	35.56	27.06	181,169	2,653	1,598	1136	133	622
	18	26.05	18.21	147,944	2,245	1,379			

Table 4.4 – Novel miRNAs in the root samples of *Arabidopsis thaliana* predicted by BrumiR.

miRID	chr:pos	mature sequence	precursor sequence
miR-1	chr5:10807602-10807763	ACCAAAAACGAAACATTCCCC	TTTATCTGTTAATTTCGTTAGGGGCAATTTTTCGTTTTTGGTGTGGGTATTTTGCATCAATTG GAGTGAGTAGAAGGAGAGGATTGATTGATTTGGTGTTCCAATTCTACCAACCGGAAAGGAT TAGAAGCGATGATGTATCTTCAGACCAACTATTACAT
miR-3	chr3:9240992-9241104	GGATGAAAGGTTTGACTAGAACT	AATAAATTGGATTTTTTAGTTAGAAAGGTTTGGCAGGACGTTATTTACTAAAAAATAAATGA GTTTTTTAGGATGAAAGGTTTGACTAGAACTGAAGATTTATGTTTATTAT
miR-8	chr1:29612248-29612361	ATTATGGACCGTCCAACTTGGCCC	TGGGCTGACCATGGACTTGCCCATATGGACATGGTCTTTATTGGGCATGGACATTTTCGGAC CATTGTCCATTATGGACCGTCCAACTTGGCCCATAAAAAAAACTGTCCGTA

Table 4.5 – Novel microRNAs and their putative interactions obtained using psRNATarget. miRNA_Acc.: microRNA identification; Target_Acc.: mRNA target identification, linked to the *Arabidopsis thaliana* mRNA library with the Araport V11 genome annotation. Expectation: mismatches penalty between mature small RNA and the target sequence, the lower the value the better the prediction (with 5.0 as a maximum threshold). Inhibition: refers to the possible mechanisms used by the sRNA to regulate its mRNA target, described in plants. Target_Desc: refers to the gene description for the mRNA target, found in the Araport V11 annotation. Multiplicity: indicates how many times a sRNA has a target sequence in a unique mRNA.

miRNA_Acc.	Target_Acc.	Expectation	Inhibition	Target_Desc.	Multiplicity	reference
miR-1	AT1G66000.1	2.0	Cleavage	hypothetical protein (DUF577)	1	1
miR-1	AT2G30700.1	2.0	Cleavage	GPI-anchored protein	1	2
miR-1	AT4G16250.1	3.0	Cleavage	phytochrome D	1	3
miR-1	AT4G24740.5	3.0	Cleavage	LAMMER-type protein kinase AFC2	1	4, 5
miR-1	AT5G03670.2	3.5	Cleavage	histone-lysine N-methyltransferase SETD1B-like protein	1	6
miR-1	AT3G04450.1	3.5	Cleavage	Homeodomain-like superfamily protein	1	7
miR-1	AT4G19920.1	3.5	Cleavage	Toll-Interleukin-Resistance (TIR) domain family protein	1	8
miR-1	AT2G10608.1	3.5	Cleavage	transmembrane protein	1	9
miR-1	AT2G32680.1	3.5	Cleavage	receptor like protein 23	1	10
miR-1	AT3G18480.1	3.5	Cleavage	CCAAT-displacement protein alternatively spliced product	1	11, 12
miR-1	AT1G14630.2	3.5	Cleavage	XRI1-like protein	1	13
miR-1	AT1G70590.1	3.5	Cleavage	F-box family protein	1	14
miR-1	AT3G21870.1	3.5	Cleavage	cyclin	1	15
miR-1	AT5G49100.1	3.5	Translation	vitellogenin-like protein	1	16
miR-3	AT5G66950.1	2.0	Cleavage	Pyridoxal phosphate (PLP)-dependent transferases superfamily protein	1	1
miR-3	AT2G18720.3	2.5	Cleavage	Translation elongation factor EF1A/initiation factor IF2gamma family protein	1	2
miR-3	AT1G56050.1	2.5	Translation	GTP-binding protein-like protein	1	3
miR-3	AT2G18720.2	2.5	Cleavage	Translation elongation factor EF1A/initiation factor IF2gamma family protein	1	2
miR-3	AT2G18720.1	2.5	Cleavage	Translation elongation factor EF1A/initiation factor IF2gamma family protein	1	2
miR-3	AT3G07540.1	2.5	Cleavage	Actin-binding FH2 (formin homology 2) family protein	1	4
miR-3	AT3G08780.1	3.0	Cleavage	BRISC complex subunit Abro1-like protein	1	5
miR-3	AT4G34920.1	3.0	Cleavage	PLC-like phosphodiesterases superfamily protein	1	6
miR-3	AT2G38060.1	3.0	Cleavage	phosphate transporter	1	7
miR-3	AT3G56080.1	3.0	Cleavage	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	1	8
miR-3	AT2G41860.1	3.0	Cleavage	calcium-dependent protein kinase 14	1	9
miR-3	AT4G17570.2	3.0	Cleavage	GATA transcription factor 26	1	10
miR-3	AT4G35620.1	3.5	Cleavage	Cyclin	1	11
miR-3	AT5G22160.1	3.5	Cleavage	transmembrane protein	1	*
miR-3	AT5G49870.2	3.5	Cleavage	Mannose-binding lectin superfamily protein	1	12
miR-3	AT2G26770.3	3.5	Cleavage	plectin-like protein	1	*
miR-8	AT4G25100.4	2.5	Cleavage	Fe superoxide dismutase 1	1	1, 2
miR-8	AT5G61630.1	3.0	Cleavage	transmembrane protein	1	3
miR-8	AT5G54020.1	3.5	Cleavage	Cysteine/Histidine-rich C1 domain family protein	2	4
miR-8	AT2G39040.1	3.5	Cleavage	Peroxidase superfamily protein	1	5, 6, 7
miR-8	AT1G04840.1	3.5	Cleavage	Tetratricopeptide repeat (TPR)-like superfamily protein	1	8
miR-8	AT3G20010.8	3.5	Cleavage	SNF2 domain-containing protein / helicase domain-containing protein / zinc finger protein-like protein	1	9, 10,11
miR-8	AT3G53570.5	3.5	Cleavage	serine/threonine-protein kinase AFC1	1	9, 10,11
miR-8	AT1G10580.1	3.5	Cleavage	Transducin/WD40 repeat-like superfamily protein	1	14, 15
miR-8	AT2G02070.2	3.5	Cleavage	indeterminate(ID)-domain 5	1	16, 17

BrumiR commands

BrumiR commands for real and simulated benchmark.

```
#trimming raw sequences using fastp
fastp --adapter_fasta  ../adapters.fa -i <prefix>.fastq.gz
                                      -o <prefix>.trim.fastq.gz


#running miRDeep2
mapper.pl <prefix>.fa -c -p genome-index -m -q -s <prefix>.reads_collapsed.fa
                         -t <prefix>.reads_collapsed_vs_genome.arf -v -o 2

miRDeep2.pl <prefix>.reads_collapsed.fa genome.fna
            <prefix>.reads_collapsed_vs_genome.arf
             none none none 2><prefix>.report.log


#running miR-PREFeR
python process-reads-fasta.py samplelist.txt <prefix>.fa <prefix>.fa

python bowtie-align-reads.py -p 2 -k 20 -f -r genome.fna <prefix>.fa.processed

python miR_PREFeR.py -L -k pipeline config.file


#running BrumiR
perl brumir.pl -a <prefix>.trim.fastq.gz -p <prefix> -T 10 -R 2 > <prefix>.log


#running BrumiR2Reference
#for animal species
perl brumir2reference.pl -a <prefix>.candidate_miRNA.fasta -b genome.fna
                         -t 4 -p <prefix>


#for plant species
perl brumir2reference.pl -a <prefix>.candidate_miRNA.fasta -b genome.fna
                         -t 4 -p <prefix> -x 1
```

4.4 Conclusion.

In the previous chapter, I had introduced the core algorithm of BrumiR and the tool miRsim to benchmark BrumiR with simulated reads. In this chapter, I presented the benchmark with real datasets and I introduced the BrumiR2reference tool. We extensively benchmarked the BrumiR toolkit on animal and plant species using simulated and real datasets. The benchmark results show that BrumiR is very sensitive, is the fastest tool, and its predictions were supported by the characteristic hairpin structure of miRNAs. Finally, we showed the power of BrumiR for discovering novel miRNAs in the model plant *Arabidopsis thaliana*. We sequenced a total of 18 sRNA-seq libraries from different stages of root development and used the BrumiR toolkit to analyze our data. We annotated three novel miRNAs involved in root development, showing in a real biological situation how BrumiR is able to infer novel information even in highly annotated genomes. Finally, we also applied BrumiR to the discovery of miRNAs of *A. thaliana* and identified three novel high-confidence miRNAs involved in root development. These putative miRNAs were not discovered before by any other software, thereby showing the potential of using different approaches even in the case where high quality genomes are available.

Conclusion and Perspectives

In the introduction of this thesis, I pointed out all the different issues to well annotate and predict miRNA candidates despite all the advances in NGS and in algorithmic tools for the discovery of miRNAs. Moreover, when we work with non-model species, being able to obtain a good and reliable miRNA prediction becomes an even greater challenge.

In Chapter 2, we worked with sRNA-seq and mRNA-seq combined from *M. hyopneumoniae* and *S. scrofa* in a host-pathogen association. As the main tools for annotation and database repositories did not have much information associated with *S. scrofa*, and the target prediction tools available were not designed for *S. scrofa*, we had to use the *Homo sapiens* model which is quite close but perhaps missed annotations that were specific of the species itself. However, we were able to identify 1,268 genes and 170 miRNAs that were significantly modified post-infection. Up-regulated mRNAs were enriched in genes related to redox homeostasis and antioxidant defense, known to be regulated by the transcription factor NRF2 in related species. Down-regulated mRNAs were enriched in genes associated with cytoskeleton and ciliary functions. The bioinformatic analyses performed suggested a correlation between changes in miRNA and mRNA levels. Indeed, we detected down-regulation of miRNAs predicted to target antioxidant genes and up-regulation of miRNAs targeting ciliary and cytoskeleton genes. Interestingly, most down-regulated miRNAs were detected in exosome-like vesicles suggesting that *M. hyopneumoniae* infection induces a modification of the composition of NPTr-released vesicles. Taken together, our data indicate that *M. hyopneumoniae* elicits an antioxidant response induced by NRF2 in infected cells. In addition, we proposed that ciliostasis caused by this pathogen is partially explained by the down-regulation of ciliary genes.

The work presented in this chapter further allowed us to identify the real challenges and difficulties in using the current state-of-the-art tools. Especially, I noticed that the main problem came right from the start, namely from the annotation of miRNAs. This step is indeed crucial for everything that comes later. I therefore decided to focus my thesis on the miRNA discovery problem and to try to contribute to improve the current methods by developing a new method based directly from the reads, independent of a reference genome or of prior knowledge, thus enabling to deal with non model species that may be completely novel in terms of characteristics and features in relation to those currently known.

In this context, I developed BrumiR, an algorithm that is able to discover miRNAs directly

and exclusively from sRNA-seq data, enabling the identification of mature miRNAs in model and non-model species with or without a reference genome, encompassing the plant and animal kingdoms. The BrumiR toolkit implements the following algorithms: 1) a new discovery miRNA tool (BRUMIR-core), 2) a specific genome mapper (BrumiR2reference), and 3) a sRNA-seq read simulator (miRsim). I introduce BrumiR and miRsim in Chapter 3, while BrumiR2reference is presented in Chapter 4.

I benchmarked BrumiR with datasets encompassing animal and plant species using simulated and real sRNA-seq experiments. I divided the benchmark of BrumiR in two parts. First, in Chapter 3 I present a benchmark on simulated reads generated by miRsim. Second, in order to test the performance of BrumiR on real data, I downloaded 20 public datasets to benchmark BrumiR; the results are presented in Chapter 4. In both cases, we showed that BrumiR is capable of identifying mature miRNAs based only on the sequence information, and that it generates results that are better or comparable to the state-of-the-art tools on simulated and real datasets. We further tested the usefulness of the BrumiR toolkit for discovering novel miRNAs potentially involved in the regulation of the root development of the extensively annotated genome of *Arabidopsis thaliana*. To achieve this, we sequenced sRNA-seq libraries from roots of *A. thaliana*, and we were able to annotate 3 novel putative miRNAs which are also presented in Chapter 4.

Unlike the state-of-the-art tools, BrumiR starts by encoding the sRNA-seq reads using a de Bruijn graph, thus avoiding a read-mapping step and the dependency on previous miRNA annotations, which makes BrumiR the fastest method and provides a stand-alone package for running locally all the analyses. It further generates an output that is compatible with the Bandage software, which can be employed to visualize and explore the results of BrumiR in a user-friendly way. In this context, the results obtained show that BrumiR reaches the highest recall for miRNA discovery, while at the same time being much faster and more efficient than the state-of-the-art tools evaluated. The latter allows BrumiR to analyze a large number of sRNA-seq experiments, from plant or animal species. Moreover, BrumiR detects additional information regarding other expressed sequences (sRNAs, isomiRs, etc.), thus maximizing the biological insight gained from sRNA-seq experiments.

Additionally, a critical step of genome-based miRNA discovery tools is to identify the precursor sequence when a reference genome is available. BrumiR introduces a new mapping approach, BrumiR2Reference, that is presented in Chapter 4. BrumiR2reference scans every possible hairpin precursor in the genome, when such is available, for all the BrumiR predictions. As the hairpin structure is determined using the predicted mature miRNA instead of the reads, this alignment can support mismatches and indels and handles the case of multi-mapped candidates (due to repetitive regions of the genome). Such features distinguish BrumiR from the current genome-based methods.

Although BrumiR shows a good performance on real and synthetic benchmarks, we can observe that the higher number of predictions it makes results in a lower accuracy in some of the data sets evaluated. This problem is improved by BrumiR2reference which refines the

BrumiR predictions, thus reducing the number of false-positive putative miRNAs but not eliminating them altogether. I thus plan in future to further reduce such false-positive rate by using a Random-Forest classifier trained on the high confidence mature sequences available in miRBase. The latter will improve the accuracy of BrumiR even in the case when a reference genome is not available.

In terms of computational resources and usability, BrumiR is the fastest method and provides a stand-alone package for running locally all the analyses. Although BrumiR-core is the fastest tool to make miRNA predictions, when we use BrumiR2reference, the exhaustive search for precursor sequences takes longer, thereby increasing the final computation time of BrumiR+BrumiR2reference. A suitable way to reduce and facilitate these computations is through parallelization of the BrumiR2reference code to distribute and run different parts of the process simultaneously depending on the computational resources available. Providing a web-server platform to run BrumiR as do several other miRNA discovery tools could also make it easier to be used by biologists who do not have as much expertise using stand-alone tools.

As part of the international collaborative project called NOIR of which we are members, we will have to identify the miRNAs and other types of sRNAs involved in the beneficial interaction between the plant *A. thaliana* and the fungi *Trichoderma atroviride*, and those involved in the host-pathogen interaction between the fish *Salmon salar* and the bacterium *Piscirickettsia salmonis*. BrumiR will be a central player to understand these interactions, and the idea then is to improve or develop entirely new features of BrumiR, in particular to be able to handle other types of sRNAs than miRNAs.

In summary, we presented in this thesis a new and versatile method that implements novel algorithmic ideas for the study of miRNAs and thus complements and extends the currently existing approaches. The code of BrumiR is freely available at (https://github.com/cmoraga/BrumiR).

Bibliography

[1] A reference standard for genome biology. *Nature Biotechnology*, 36(12):1121, December 2018.

[2] MN Abu-Zahr and M Butler. Growth, cytopathogenicity and morphology of *Mycoplasma gallisepticum* and *M. gallinarum* in tracheal explants. *Journal of Comparative Pathology*, 86(3):455–463, 1976.

[3] Ludmila Alekseeva, Lucie Rault, Sintia Almeida, Patrick Legembre, Valérie Edmond, Vasco Azevedo, Anderson Miyoshi, Sergine Even, Frédéric Taieb, Yannick Arlot-Bonnemains, et al. *Staphylococcus aureus*-induced g2/m phase transition delay in host epithelial cells increases bacterial infective efficiency. *PloS One*, 8(5):e63279, 2013.

[4] Margaret Alexander, Ruozhen Hu, Marah C Runtsch, Dominique A Kagele, Timothy L Mosbruger, Tanya Tolmachova, Miguel C Seabra, June L Round, Diane M Ward, and Ryan M O'Connell. Exosome-delivered microRNAs modulate the inflammatory response to endotoxin. *Nature Communications*, 6:7321, 2015.

[5] Begum Alural, Aysegul Ozerdem, Jens Allmer, Kursad Genc, and Sermin Genc. Lithium protects against paraquat neurotoxicity by NRF2 activation and miR-34a inhibition in SH-SY5Y cells. *Frontiers in Cellular Neuroscience*, 9:209, 2015.

[6] Clara Amid, Blaise TF Alako, Vishnukumar Balavenkataraman Kadhirvelu, Tony Burdett, Josephine Burgin, Jun Fan, Peter W Harrison, Sam Holt, Abdulrahman Hussein, Eugene Ivanov, et al. The European Nucleotide Archive in 2019. *Nucleic Acids Research*, 48(D1):D70–D76, 2020.

[7] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq,Äîa python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

[8] Ernesto Aparicio-Puerta, Ricardo Lebrén, Antonio Rueda, Cristina Gómez-Martín, Stavros Giannoukakos, David Jaspez, José María Medina, Andreja Zubkovic, Igor Jurak, Bastian Fromm, Juan Antonio Marchal, José Oliver, and Michael Hackenberg. sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Research*, 47(W1):W530–W535, 2019.

[9] Minako Araake. Comparison of ciliostasis by mycoplasmas in mouse and chicken tracheal organ cultures. *Microbiology and Immunology*, 26(1):1–14, 1982.

[10] Sathya N Kulappu Arachchige, Neil D Young, Pollob K Shil, Alistair R Legione, Anna Kanci Condello, Glenn F Browning, and Nadeeka K Wawegama. Differential Response of the Chicken Trachea to Chronic Infection with Virulent *Mycoplasma gallisepticum* Strain Ap3AS and Vaxsafe MG (Strain ts-304): a Transcriptional Profile. *Infection and Immunity*, 88(5), 2020.

[11] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[12] Janhavi Athale, Allison Ulrich, Nancy Chou MacGarvey, Raquel R Bartz, Karen E Welty-Wolf, Hagir B Suliman, and Claude A Piantadosi. Nrf2 promotes alveolar mitochondrial biogenesis and resolution of lung injury in staphylococcus aureus pneumonia in mice. *Free Radical Biology and Medicine*, 53(8):1584–1594, 2012.

[13] Michael J. Axtell and Blake C. Meyers. Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. *The Plant Cell*, 30(2):272–284, 2018.

[14] Hua Bao, Arun Kommadath, Guanxiang Liang, Xu Sun, Adriano S Arantes, Christopher K Tuggle, Shawn MD Bearson, Graham S Plastow, Paul Stothard, et al. Genome-wide whole blood micrornaome and transcriptome analyses reveal mirna-mrna regulated host response to foodborne pathogen *Salmonella* infection in swine. *Scientific reports*, 5:12620, 2015.

[15] David P. Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[16] David P Bartel. Metazoan micrornas. *Cell*, 173(1):20–51, 2018.

[17] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[18] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8):R90, 2010.

[19] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, 2009.

[20] Daniela F Bischof, Carole Janis, Edy M Vilei, Giuseppe Bertoni, and Joachim Frey. Cytotoxicity of *Mycoplasma mycoides* subsp. mycoides small colony type to bovine epithelial cells. *Infection and Immunity*, 76(1):263–269, 2008.

[21] Shyam Biswal, Rajesh K Thimmulappa, and Christopher J Harvey. Experimental therapeutics of nrf2 as a target for prevention of bacterial exacerbations in copd. *Proceedings of the American Thoracic Society*, 9(2):47–51, 2012.

[22] Victor M Bolanos-Garcia and Tom L Blundell. Bub1 and bubr1: multifaceted kinases of the cell cycle. *Trends in biochemical sciences*, 36(3):141–150, 2011.

[23] Michele Bortolomeazzi, Enrico Gaffo, and Stefania Bortoluzzi. A survey of software tools for microRNA discovery and characterization using RNA-seq. *Briefings in Bioinformatics*, 20(3):918–930, 2019.

[24] Eleanor J. Brant and Hikmet Budak. Plant Small Non-coding RNAs and Their Roles in Biotic Stresses. *Frontiers in Plant Science*, 9, 2018.

[25] Natalie W. Breakfield, David L. Corcoran, Jalean J. Petricka, Jeffrey Shen, Juthamas Sae-Seaw, Ignacio Rubio-Somoza, Detlef Weigel, Uwe Ohler, and Philip N. Benfey. High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in *Arabidopsis*. *Genome Research*, 2011.

[26] Kamil Brzóska, Tomasz M Stępkowski, and Marcin Kruszewski. Basal pir expression in hela cells is driven by nrf2 via evolutionary conserved antioxidant response element. *Molecular and cellular biochemistry*, 389(1-2):99–111, 2014.

[27] Hikmet Budak, Melda Kantar, Reyyan Bulut, and Bala Ani Akpinar. Stress responsive miRNAs and isomiRs in cereals. *Plant Science*, 235:1–13, June 2015.

[28] Tracey A Burnett, Katrin Dinkla, Manfred Rohde, Gursharan S Chhatwal, Cord Uphoff, Mukesh Srivastava, Stuart J Cordwell, Steven Geary, Xiaofen Liao, F Chris Minion, et al. P159 is a proteolytically processed, surface adhesin of *Mycoplasma hyopneumoniae*: defined domains of p159 bind heparin and promote adherence to eukaryote cells. *Molecular microbiology*, 60(3):669–686, 2006.

[29] Andrew Chatr-Aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, et al. The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379, 2017.

[30] R Chaudhry, A Ghosh, A Chandolia, et al. Pathogenesis of *Mycoplasma pneumoniae*: An update. *Indian Journal of Medical Microbiology*, 34(1):7, 2016.

[31] Li Chen, Alyssa Charrier, Yu Zhou, Ruju Chen, Bo Yu, Kitty Agarwal, Hidekazu Tsukamoto, L James Lee, Michael E Paulaitis, and David R Brigstock. Epigenetic regulation of connective tissue growth factor by microrna-214 delivery in exosomes from mouse or human hepatic stellate cells. *Hepatology*, 59(3):1118–1129, 2014.

[32] Liang Chen, Liisa Heikkinen, Changliang Wang, Yang Yang, Huiyan Sun, and Garry Wong. Trends in the development of miRNA bioinformatics tools. *Briefings in Bioinformatics*, 20(5):1836–1852, 2019.

[33] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018. Publisher: Oxford Academic.

[34] Ying Cheng, Wenhua Kuang, Yongchang Hao, Donglin Zhang, Ming Lei, Li Du, Hanwei Jiao, Xiaoru Zhang, and Fengyang Wang. Downregulation of mir-27a* and mir-532-5p and upregulation of mir-146a and mir-155 in lps-induced raw264. 7 macrophage cells. *Inflammation*, 35(4):1308–1313, 2012.

[35] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32(12):i201–i208, 2016.

[36] Rayan Chikhi and Guillaume Rizk. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for Molecular Biology*, 8(1):22, 2013.

[37] Hye-Youn Cho, Sekhar P Reddy, and Steven R Kleeberger. Nrf2 defends the lung from oxidative stress. *Antioxidants & redox signaling*, 8(1-2):76–87, 2006.

[38] Raymond J Cho, Mingxia Huang, Michael J Campbell, Helin Dong, Lars Steinmetz, Lisa Sapinoso, Garret Hampton, Stephen J Elledge, Ronald W Davis, and David J Lockhart. Transcriptional regulation and function during the human cell cycle. *Nature genetics*, 27(1):48, 2001.

[39] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. Why are de Bruijn graphs useful for genome assembly? *Nature biotechnology*, 29(11):987, 2011.

[40] Li-Li Dai, Jin-Xia Gao, Cheng-Gang Zou, Yi-Cheng Ma, and Ke-Qin Zhang. mir-233 modulates the unfolded protein response in *C. elegans* during *Pseudomonas aeruginosa* infection. *PLoS pathogens*, 11(1):e1004606, 2015.

[41] Xinbin Dai, Zhaohong Zhuang, and Patrick Xuechun Zhao. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Research*, 46(W1):W49–W54, 2018.

[42] Padideh Danaee, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, and David Hendrix. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research*, 46(11):5381–5394, 2018.

[43] Mary C DeBey and Richard F Ross. Ciliostasis and loss of cilia induced by *Mycoplasma hyopneumoniae* in porcine tracheal organ cultures. *Infection and immunity*, 62(12):5312–5318, 1994.

[44] Sebastian Deorowicz, Agnieszka Debudaj-Grabysz, and Szymon Grabowski. Disk-based k-mer counting on a PC. *BMC Bioinformatics*, 14(1):160, 2013.

[45] Tamsyn Derrick, Anna R Last, Sarah E Burr, Meno Nabicassa, Eunice Cassama, Robin L Bailey, David CW Mabey, Matthew J Burton, Martin J Holland, et al. Inverse relationship between microrna-155 and-184 expression with increasing conjunctival inflammation during ocular *Chlamydia trachomatis* infection. *BMC infectious diseases*, 16(1):60, 2015.

[46] Tamsyn Derrick, Megha Rajasekhar, Sarah E Burr, Hassan Joof, Pateh Makalo, Robin L Bailey, David CW Mabey, Matthew J Burton, Martin J Holland, et al. Conjunctival microrna expression in inflammatory trachomatous scarring. *PLoS neglected tropical diseases*, 7(3):e2117, 2013.

[47] Tamsyn Derrick, Athumani M Ramadhani, Karim Mtengai, Patrick Massae, Matthew J Burton, and Martin J Holland. mirnas that associate with conjunctival inflammation and ocular *Chlamydia trachomatis* infection do not predict progressive disease. *Pathogens and disease*, 75(2), 2017.

[48] Steven P Djordjevic, Stuart J Cordwell, Michael A Djordjevic, Jody Wilton, and F Chris Minion. Proteolytic processing of the *Mycoplasma hyopneumoniae* cilium adhesin. *Infection and immunity*, 72(5):2791–2802, 2004.

[49] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[50] Dilip A. Durai and Marcel H. Schulz. Informed kmer selection for *de novo* transcriptome assembly. *Bioinformatics*, 32(11):1670–1677, 2016.

[51] Mélodie Duval, Pascale Cossart, and Alice Lebreton. Mammalian micrornas and long noncoding rnas in the host-bacterial pathogen crosstalk. In *Seminars in cell & developmental biology*, volume 65, pages 11–19. Elsevier, 2017.

[52] Petr Dvorak, Yuliya Krasylenko, Miroslav Ovecka, Jasim Basheer, Veronika Zapletalová, Jozef Samaj, and Tomás Takakç. FSD1: developmentally-regulated plastidial, nuclear

and cytoplasmic enzyme with anti-oxidative and osmoprotective role. *Plant, Cell & Environment*, 2020.

[53] Noah Fahlgren, Miya D. Howell, Kristin D. Kasschau, Elisabeth J. Chapman, Christopher M. Sullivan, Jason S. Cumbie, Scott A. Givan, Theresa F. Law, Sarah R. Grant, Jeffery L. Dangl, and James C. Carrington. High-Throughput Sequencing of *Arabidopsis* microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. *PLOS ONE*, 2(2):e219, 2007.

[54] Lina Fassi Fehri, Manuel Koch, Elena Belogolova, Hany Khalil, Christian Bolz, Behnam Kalali, Hans J. Mollenkopf, Macarena Beigier-Bompadre, Alexander Karlas, Thomas Schneider, Yuri Churin, Markus Gerhard, and Thomas F. Meyer. *Helicobacter pylori* Induces miR-155 in T Cells in a cAMP-Foxp3-Dependent Manner. *PLoS One*, 5(3), March 2010.

[55] R. Feinbaum and V. Ambros. The timing of lin-4 RNA accumulation controls the timing of postembryonic developmental events in *Caenorhabditis elegans*. *Developmental Biology*, 210(1):87–95, June 1999.

[56] Jianxing Feng, Clifford A Meyer, Qian Wang, Jun S Liu, X Shirley Liu, and Yong Zhang. Gfold: a generalized fold change for ranking differentially expressed genes from rna-seq data. *Bioinformatics*, 28(21):2782–2788, 2012.

[57] Angeles Fernandez-Gonzalez, Stella Kourembanas, Todd A Wyatt, and S Alex Mitsialis. Mutation of murine adenylate kinase 7 underlies a primary ciliary dyskinesia phenotype. *American journal of respiratory cell and molecular biology*, 40(3):305–313, 2009.

[58] Mariana Galvão Ferrarini, Scheila Gabriele Mucha, Delphine Parrot, Guillaume Meiffrein, Jose Fernando Ruggiero Bachega, Gilles Comte, Arnaldo Zaha, and Marie-France Sagot. Hydrogen peroxide production and myo-inositol metabolism as important traits for virulence of *Mycoplasma hyopneumoniae*. *Molecular microbiology*, 108(6):683–696, 2018.

[59] Alexis Forterre, Audrey Jalabert, Karim Chikh, Sandra Pesenti, Vanessa Euthine, Aurélie Granjon, Elizabeth Errazuriz, Etienne Lefai, Hubert Vidal, and Sophie Rome. Myotube-derived exosomal mirnas downregulate sirtuin1 in myoblasts during muscle cell differentiation. *Cell cycle*, 13(1):78–89, 2014.

[60] Simon Fourquet, Raphaël Guerois, Denis Biard, and Michel B Toledano. Activation of nrf2 by nitrosative agents and h2o2 involves keap1 disulfide formation. *Journal of Biological Chemistry*, 285(11):8463–8471, 2010.

[61] Marc R Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, 26(4):407–415, 2008.

[62] Marc R Friedländer, Sebastian D Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. mirdeep2 accurately identifies known and hundreds of novel microrna genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52, 2011.

[63] Marc R. Friedländer, Sebastian D. Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52, 2012.

[64] N. E. Fusenig, D. Breitkreutz, P. Boukamp, P. Tomakidi, and H. J. Stark. Differentiation and tumor progression. *Recent Results in Cancer Research. Fortschritte Der Krebsforschung. Progres Dans Les Recherches Sur Le Cancer*, 139:1–19, 1995.

[65] Ai-Mei Gao, Xiao-Yu Zhang, and Zun-Ping Ke. Apigenin sensitizes bel-7402/adm cells to doxorubicin through inhibiting mir-101/nrf2 pathway. *Oncotarget*, 8(47):82085, 2017.

[66] Jorge A Girón, Monika Lange, and Joel B Baseman. Adherence, fibronectin binding, and induction of cytoskeleton reorganization in cultured human cells by *Mycoplasma penetrans*. *Infection and Immunity*, 64(1):197–208, 1996.

[67] John C Gomez, Hong Dang, Jessica R Martin, and Claire M Doerschuk. Nrf2 modulates host defense during *Streptococcus pneumoniae* pneumonia in mice. *The Journal of Immunology*, 197(7):2864–2879, 2016.

[68] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, 29(7):644, 2011.

[69] Daniel B Graham, Guadalupe J Jasso, Amanda Mok, Gautam Goel, Aylwin CY Ng, Raivo Kolde, Mukund Varma, John G Doench, David E Root, Clary B Clish, et al. Nitric oxide engages an anti-inflammatory feedback loop mediated by peroxiredoxin 5 in phagocytes. *Cell reports*, 24(4):838–850, 2018.

[70] Sam Griffiths-Jones. The microrna registry. *Nucleic Acids Research*, 32(suppl_1):D109–D111, 2004.

[71] Helge Grosshans and Witold Filipowicz. The expanding world of small RNAs. *Nature*, 451(7177):414–416, January 2008. Number: 7177 Publisher: Nature Publishing Group.

[72] Adam Grundhoff and Christopher S. Sullivan. Virus-encoded microRNAs. *Virology*, 411(2):325–343, March 2011.

[73] Qingmei Guan, Jianmin Wu, Xiule Yue, Yanyan Zhang, and Jianhua Zhu. A nuclear calcium-sensing pathway is critical for gene regulation and salt stress tolerance in *Arabidopsis*. *PLoS Genet*, 9(8):e1003755, 2013.

[74] Sven Halbedel, Claudine Hames, and Jörg Stülke. In vivo activity of enzymatic and regulatory components of the phosphoenolpyruvate: sugar phosphotransferase system in *Mycoplasma pneumoniae*. *Journal of Bacteriology*, 186(23):7936–7943, 2004.

[75] Claudine Hames, Sven Halbedel, Michael Hoppert, Joachim Frey, and Jörg Stülke. Glycerol metabolism is important for cytotoxicity of *Mycoplasma pneumoniae*. *Journal of Bacteriology*, 191(3):747–753, 2009.

[76] John D Hayes and Albena T Dinkova-Kostova. The nrf2 regulatory network provides an interface between redox and intermediary metabolism. *Trends in Biochemical Sciences*, 39(4):199–218, 2014.

[77] John D Hayes and Michael McMahon. Nrf2 and keap1 mutations: permanent activation of an adaptive response in cancer. *Trends in Biochemical Sciences*, 34(4):176–188, 2009.

[78] Peter S Hegan, Eric Ostertag, Aron M Geurts, and Mark S Mooseker. Myosin id is required for planar cell polarity in ciliated tracheal and ependymal epithelial cells. *Cytoskeleton*, 72(10):503–516, 2015.

[79] Susan Higashi, Cyril Fournier, Christian Gautier, Christine Gaspin, and Marie-France Sagot. Mirinho: An efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data. *BMC Bioinformatics*, 16(1):179, 2015.

[80] Jun Hong, Ying Wang, Bang-Chuan Hu, Liang Xu, Jing-Quan Liu, Min-Hua Chen, Jin-Zhu Wang, Fang Han, Yang Zheng, Xu Chen, et al. Transcriptional downregulation of microrna-19a by ros production and nf-κb deactivation governs resistance to oxidative stress-initiated apoptosis. *Oncotarget*, 8(41):70967, 2017.

[81] Hristo B. Houbaviy, Michael F. Murray, and Phillip A. Sharp. Embryonic stem cell-specific MicroRNAs. *Developmental Cell*, 5(2):351–358, August 2003.

[82] Li-Ching Hsieh, Shu-I. Lin, Arthur Chun-Chieh Shih, June-Wei Chen, Wei-Yi Lin, Ching-Ying Tseng, Wen-Hsiung Li, and Tzyy-Jen Chiou. Uncovering small RNA-mediated responses to phosphate deficiency in *Arabidopsis* by deep sequencing. *Plant Physiology*, 151(4):2120–2132, 2009.

[83] Yasuko Inaba, Kyosuke Shinohara, Yanick Botilde, Ryo Nabeshima, Katsuyoshi Takaoka, Rieko Ajima, Lynda Lamri, Hiroyuki Takeda, Yumiko Saga, Tetsuya Nakamura, et al. Transport of the outer dynein arm complex to cilia requires a cytoplasmic protein lrrc6. *Genes to Cells*, 21(7):728–739, 2016.

[84] Ashwani Jha and Ravi Shankar. miReader: Discovering Novel miRNAs in Species without Sequenced Genome. *PLOS ONE*, 8(6):e66857, 2013.

[85] Allison Jones, Ann-Beth Jonsson, and Helena Aro. *Neisseria gonorrhoeae* infection causes a g1 arrest in human epithelial cells. *The FASEB Journal*, 21(2):345–355, 2007.

[86] Ioanna Kalvari, Joanna Argasinska, Natalia Quinones-Olvera, Eric P. Nawrocki, Elena Rivas, Sean R. Eddy, Alex Bateman, Robert D. Finn, and Anton I. Petrov. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, 46(D1):D335–D342, 2018.

[87] Ioanna Kalvari, Eric P. Nawrocki, Joanna Argasinska, Natalia Quinones-Olvera, Robert D. Finn, Alex Bateman, and Anton I. Petrov. Non-Coding RNA Analysis Using the Rfam Database. *Current Protocols in Bioinformatics*, 62(1):e51, 2018.

[88] Moon-Il Kang, Akira Kobayashi, Nobunao Wakabayashi, Sang-Geon Kim, and Masayuki Yamamoto. Scaffolding of keap1 to the actin cytoskeleton controls the function of nrf2 as key regulator of cytoprotective phase 2 genes. *Proceedings of the National Academy of Sciences*, 101(7):2046–2051, 2004.

[89] Thomas W Kensler, Nobunao Wakabayashi, and Shyam Biswal. Cell survival responses to environmental stresses via the keap1-nrf2-are pathway. *Annu. Rev. Pharmacol. Toxicol.*, 47:89–116, 2007.

[90] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microrna target recognition. *Nature Genetics*, 39(10):1278, 2007.

[91] Kotaro Kiga, Hitomi Mimuro, Masato Suzuki, Aya Shinozaki-Ushiku, Taira Kobayashi, Takahito Sanada, Minsoo Kim, Michinaga Ogawa, Yuka W. Iwasaki, Hiroyuki Kayo, Yoko Fukuda-Yuzawa, Masakazu Yashiro, Masashi Fukayama, Taro Fukao, and Chihiro Sasakawa. Epigenetic silencing of miR-210 increases the proliferation of gastric epithelium during chronic *Helicobacter pylori* infection. *Nature Communications*, 5(1):4497, September 2014. Number: 1 Publisher: Nature Publishing Group.

[92] V Narry Kim and Jin-Wu Nam. Genomics of microrna. *Trends in Genetics*, 22(3):165–173, 2006.

[93] Yong Ha Kim, Jun Yeol Choi, Yeontae Jeong, Debra J Wolgemuth, and Kunsoo Rhee. Nek2 localizes to multiple sites in mitotic cells, suggesting its involvement in multiple cellular functions during the cell cycle. *Biochemical and Biophysical Research communications*, 290(2):730–736, 2002.

[94] Daisuke Kobayashi and Hiroyuki Takeda. Ciliary motility: the components and cytoplasmic preassembly mechanisms of the axonemal dyneins. *Differentiation*, 83(2):S23–S29, 2012.

[95] Marek Kokot, Maciej Dlugosz, and Sebastian Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics (Oxford, England)*, 33(17):2759–2761, 2017.

[96] Ana Kozomara and Sam Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73, 2013.

[97] Ana Kozomara and Sam Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(Database issue):D68–D73, 2014.

[98] Jan Krüger and Marc Rehmsmeier. Rnahybrid: microrna target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(suppl_2):W451–W454, 2006.

[99] W. Y. Kuo, C. H. Huang, A. C. Liu, C. P. Cheng, S. H. Li, W. C. Chang, C. Weiss, A. Azem, and T. L. Jinn. Chaperonin 20 mediates iron superoxide dismutase (FeSOD) activity independent of its co-chaperonin role in *Arabidopsis* chloroplasts. *The New Phytologist*, 197(1):99–110, 2013.

[100] Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, and Thomas Tuschl. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543):853–858, 2001.

[101] Marine Lambert, Abderrahim Benmoussa, and Patrick Provost. Small Non-Coding RNAs Derived from Eukaryotic Ribosomal RNA. *Non-Coding RNA*, 5(1):16, 2019.

[102] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[103] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[104] Nelson C. Lau, Lee P. Lim, Earl G. Weinstein, and David P. Bartel. An Abundant Class of Tiny RNAs with Probable Regulatory Roles in *Caenorhabditis elegans*. *Science*, 294(5543):858–862, 2001.

[105] Nathan Lawless, Amir BK Foroushani, Matthew S McCabe, Cliona O'Farrelly, and David J Lynn. Next generation sequencing reveals the expression of a unique mirna profile in response to a gram-positive bacterial infection. *PLoS One*, 8(3):e57543, 2013.

[106] Minh T. N. Le, Huangming Xie, Beiyan Zhou, Poh Hui Chia, Pamela Rizk, Moonkyoung Um, Gerald Udolph, Henry Yang, Bing Lim, and Harvey F. Lodish. MicroRNA-125b promotes neuronal differentiation in human cells by repressing multiple targets. *Molecular and Cellular Biology*, 29(19):5290–5305, October 2009.

[107] Rosalind C. Lee and Victor Ambros. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science (New York, N.Y.)*, 294(5543):862–864, 2001.

[108] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993.

[109] Yoontae Lee, Kipyoung Jeon, Jun-Tae Lee, Sunyoung Kim, and Narry Kim. MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO Journal*, 21(17):4663–4670, September 2002.

[110] Jikai Lei and Yanni Sun. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics (Oxford, England)*, 30(19):2837–2839, 2014.

[111] Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, 2003.

[112] Harry R. Lewis and Christos H. Papadimitriou. Symmetric space-bounded computation. *Theoretical Computer Science*, 19(2):161–187, 1982.

[113] Dan Li, Shuren Ma, and Elizabeth M Ellis. Nrf2-mediated adaptive response to methyl glyoxal in hepg2 cells involves the induction of akr7a2. *Chemico-biological Interactions*, 234:366–371, 2015.

[114] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[115] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[116] Na Li, Xiang Xu, Bin Xiao, En-Dong Zhu, Bo-sheng Li, Zhen Liu, Bin Tang, Quan-Ming Zou, Hua-Ping Liang, and Xu-Hu Mao. *H. pylori* related proinflammatory cytokines contribute to the induction of mir-146a in human gastric epithelial cells. *Molecular Biology Reports*, 39(4):4655–4661, 2012.

[117] Shengjun Li, Claudia Castillo-González, Bin Yu, and Xiuren Zhang. The functions of plant small RNAs in development and in stress responses. *The Plant Journal*, 90(4):654–670, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.13444.

[118] Yue Li, Zhuo Zhang, Feng Liu, Wanwipa Vongsangnak, Qing Jing, and Bairong Shen. Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic acids research*, 40(10):4298–4305, 2012.

[119] Guanxiang Liang, Nilusha Malmuthuge, Yongjuan Guan, Yuwei Ren, Philip J Griebel, et al. Altered microrna expression and pre-mrna splicing events reveal new mechanisms associated with early stage *Mycobacterium avium* subspecies *paratuberculosis* infection. *Scientific Reports*, 6:24964, 2016.

[120] Lee P Lim, Nelson C Lau, Philip Garrett-Engele, Andrew Grimson, Janell M Schelter, John Castle, David P Bartel, Peter S Linsley, and Jason M Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769, 2005.

[121] Runmao Lin, Liye He, Jiayu He, Peigang Qin, Yanran Wang, Qiming Deng, Xiaoting Yang, Shuangcheng Li, Shiquan Wang, Wenming Wang, Huainian Liu, Ping Li, and Aiping Zheng. Comprehensive analysis of microRNA-Seq and target mRNAs of rice sheath blight pathogen provides new insights into pathogenic regulatory mechanisms. *DNA Research*, 23(5):415–425, 2016.

[122] H Liu, HY Wu, WY Wang, ZL Zhao, XY Liu, and LY Wang. Regulation of mir-92a on vascular endothelial aging via mediating nrf2-keap1-are signal pathway. *Eur Rev Med Pharmacol Sci*, 21(11):2734–2742, 2017.

[123] Zhen Liu, Di Wang, Yongliang Hu, Guoyong Zhou, Chaohui Zhu, Qi Yu, Yingchun Chi, Yingli Cao, Chiyu Jia, and Quanming Zou. Microrna-146a negatively regulates ptgs2 expression induced by helicobacter pylori in human gastric epithelial cells. *Journal of Gastroenterology*, 48(1):86–92, 2013.

[124] Zhen Liu, Bin Xiao, Bin Tang, Bosheng Li, Na Li, Endong Zhu, Gang Guo, Jiang Gu, Yuan Zhuang, Xiaofei Liu, et al. Up-regulated microrna-146a negatively modulate *Helicobacter pylori*-induced inflammatory response in human gastric epithelial cells. *Microbes and Infection*, 12(11):854–863, 2010.

[125] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[126] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[127] Cheng Lu, Blake C. Meyers, and Pamela J. Green. Construction of small RNA cDNA libraries for deep sequencing. *Methods (San Diego, Calif.)*, 43(2):110–117, October 2007.

[128] Elsebet Lund, Stephan Güttinger, Angelo Calado, James E. Dahlberg, and Ulrike Kutay. Nuclear export of microRNA precursors. *Science (New York, N.Y.)*, 303(5654):95–98, January 2004.

[129] Yi Luo, Pengjun Wang, Xun Wang, Yuhao Wang, Zhiping Mu, Qingzhi Li, Yuhua Fu, Juan Xiao, Guojun Li, Yao Ma, et al. Detection of dietetically absorbed maize-derived micrornas in pigs. *Scientific reports*, 7(1):645, 2017.

[130] Qiang Ma. Role of nrf2 in oxidative stress and toxicity. *Annual Review of Pharmacology and Toxicology*, 53:401–426, 2013.

[131] Gary J Mack and Duane A Compton. Analysis of mitotic microtubule-associated proteins using mass spectrometry identifies astrin, a spindle-associated protein. *Proceedings of the National Academy of Sciences*, 98(25):14434–14439, 2001.

[132] Dominiek Maes, M Sibila, Peter Kuhnert, J Segalés, Freddy Haesebrouck, and Maria Pieters. Update on *Mycoplasma hyopneumoniae* infections in pigs: Knowledge gaps for improved disease control. *Transboundary and emerging diseases*, 65:110–124, 2018.

[133] Dominiek Maes, Marc Verdonck, Hubert Deluyker, and Aart de Kruif. Enzootic pneumonia in pigs. *Veterinary Quarterly*, 18(3):104–109, 1996.

[134] Zhumur Mallick, Bibekanand and Ghosh. *Regulatory RNAs: Basics, Methods and Applications*. Springer Science, 2012.

[135] Daniel Mapleson, Simon Moxon, Tamas Dalmay, and Vincent Moulton. MirPlex: a tool for identifying miRNAs in high-throughput sRNA datasets without a genome. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, 320(1):47–56, 2013.

[136] Olivier Marchès, Terence Neil Ledger, Michèle Boury, Masaru Ohara, Xuanlin Tu, Frédéric Goffaux, Jacques Mainil, Ilan Rosenshine, Motoyuki Sugai, Jean De Rycke, et al. Enteropathogenic and enterohaemorrhagic *Escherichia coli* deliver a novel effector called cif, which blocks cell cycle g2/m transition. *Molecular Microbiology*, 50(5):1553–1567, 2003.

[137] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17(1):10–12, 2011.

[138] Paolo Martini, Gabriele Sales, Mattia Brugiolo, Alessandro Gandaglia, Filippo Naso, Cristiano De Pittà, Michele Spina, Gino Gerosa, Francesco Chemello, Chiara Romualdi, et al. Tissue-specific expression and regulatory networks of pig micrornaome. *PLoS One*, 9(4):e89755, 2014.

[139] Anthony Mathelier and Alessandra Carbone. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, 26(18):2226–2234, 2010.

[140] Daniel R Matson and P Todd Stukenberg. Cenp-i and aurora b act as a molecular switch that ties rzz/mad1 recruitment to kinetochore attachment status. *J Cell Biol*, 205(4):541–554, 2014.

[141] Kayoko Matsushima, Hajime Isomoto, Naoki Inoue, Toshiyuki Nakayama, Tomayoshi Hayashi, Masaaki Nakayama, Kazuhiko Nakao, Toshiya Hirayama, and Shigeru Kohno. Microrna signatures in *Helicobacter pylori*-infected gastric mucosa. *International Journal of Cancer*, 128(2):361–370, 2011.

[142] Masashi Matsuyama, Andrew J Martins, Shamira Shallom, Olena Kamenyeva, Anuj Kashyap, Elizabeth P Sampaio, Juraj Kabat, Kenneth N Olivier, Adrian M Zelazny, John S Tsang, et al. Transcriptional response of respiratory epithelium to nontuberculous mycobacteria. *American Journal of Respiratory Cell and Molecular Biology*, 58(2):241–252, 2018.

[143] Claire Maudet, Miguel Mano, and Ana Eulalio. Micrornas in the interaction between host and bacterial pathogens. *FEBS Letters*, 588(22):4140–4147, 2014.

[144] Helen L May-Simera, Jessica D Gumerson, Chun Gao, Maria Campos, Stephanie M Cologna, Tina Beyer, Karsten Boldt, Koray D Kaya, Nisha Patel, Friedrich Kretschmer, et al. Loss of macf1 abolishes ciliogenesis and disrupts apicobasal polarity establishment in the retina. *Cell Reports*, 17(5):1399–1413, 2016.

[145] Blake C. Meyers, Michael J. Axtell, Bonnie Bartel, David P. Bartel, David Baulcombe, John L. Bowman, Xiaofeng Cao, James C. Carrington, Xuemei Chen, Pamela J. Green, Sam Griffiths-Jones, Steven E. Jacobsen, Allison C. Mallory, Robert A. Martienssen, R. Scott Poethig, Yijun Qi, Herve Vaucheret, Olivier Voinnet, Yuichiro Watanabe, Detlef Weigel, and Jian-Kang Zhu. Criteria for Annotation of Plant MicroRNAs. *The Plant Cell*, 20(12):3186–3190, 2008.

[146] Javier Milara, Miguel Armengot, Manuel Mata, Esteban J Morcillo, and Julio Cortijo. Role of adenylate kinase type 7 expression on cilia motility: possible link in primary ciliary dyskinesia. *American Journal of Rhinology & Allergy*, 24(3):181–185, 2010.

[147] Anthony A. Millar and Peter M. Waterhouse. Plant and animal microRNAs: similarities and differences. *Functional & Integrative Genomics*, 5(3):129–135, July 2005.

[148] Mary Mirvis, Tim Stearns, and W James Nelson. Cilium structure, assembly, and disassembly regulated by the cytoskeleton. *Biochemical Journal*, 475(14):2329–2353, 2018.

[149] Prasun J. Mishra and Glenn Merlino. MicroRNA reexpression as differentiation therapy in cancer. *The Journal of Clinical Investigation*, 119(8):2119–2123, August 2009.

[150] Dov Moldovan, Andrew Spriggs, Jun Yang, Barry J. Pogson, Elizabeth S. Dennis, and Iain W. Wilson. Hypoxia-responsive microRNAs and trans-acting small interfering RNAs in *Arabidopsis*. *Journal of Experimental Botany*, 61(1):165–177, 2010.

[151] Ryan D Morin, Michael D O'Connor, Malachi Griffith, Florian Kuchenbauer, Allen Delaney, Anna-Liisa Prabhu, Yongjun Zhao, Helen McDonald, Thomas Zeng, Martin Hirst, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome research*, 18(4):610–621, 2008.

[152] Scheila G. Mucha, Mariana G. Ferrarini, Carol Moraga, Alex Di Genova, Laurent Guyon, Florence Tardy, Sophie Rome, Marie-France Sagot, and Arnaldo Zaha. Mycoplasma hyopneumoniae J elicits an antioxidant response and decreases the expression of ciliary genes in infected swine epithelial cells. *Scientific Reports*, 10(1):13707, August 2020. Number: 1 Publisher: Nature Publishing Group.

[153] Yoshihiro Muneta, Yu Minagawa, Yoshihiro Shimoji, Yohsuke Ogawa, Hirokazu Hikono, and Yasuyuki Mori. Immune response of gnotobiotic piglets against *Mycoplasma hyopneumoniae*. *Journal of Veterinary Medical Science*, 70(10):1065–1070, 2008.

[154] Sandra M Muxel, Maria F Laranjeira-Silva, Ricardo A Zampieri, and Lucile M Floeter-Winter. *Leishmania amazonensis* induces macrophage mir-294 and mir-721 expression and modulates infection by targeting nos2 and l-arginine metabolism. *Scientific Reports*, 7:44141, 2017.

[155] Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21(suppl_-2):ii79–ii85, 2005.

[156] Afsar R Naqvi, Jezrom B Fordham, and Salvador Nares. mir-24, mir-30b, and mir-142-3p regulate phagocytosis in myeloid inflammatory cells. *The Journal of Immunology*, 194(4):1916–1927, 2015.

[157] Paul Nioi, Michael McMahon, Ken Itoh, Masayuki Yamamoto, and John D Hayes. Identification of a novel nrf2-regulated antioxidant response element (are) in the mouse nad (p) h: quinone oxidoreductase 1 gene: reassessment of the are consensus sequence. *Biochemical Journal*, 374(2):337–348, 2003.

[158] Jennifer M Noto, M Blanca Piazuelo, Rupesh Chaturvedi, Courtney A Bartel, Elizabeth J Thatcher, Alberto Delgado, Judith Romero-Gallo, Keith T Wilson, Pelayo Correa, James G Patton, et al. Strain-specific suppression of microrna-320 by carcinogenic helicobacter pylori promotes expression of the antiapoptotic protein mcl-1. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 305(11):G786–G796, 2013.

[159] Amy E. Pasquinelli, Brenda J. Reinhart, Frank Slack, Mark Q. Martindale, Mitzi I. Kuroda, Betsy Maller, David C. Hayward, Eldon E. Ball, Bernard Degnan, Peter Müller,

Jurg Spring, Ashok Srinivasan, Mark Fishman, John Finnerty, Joseph Corbo, Michael Levine, Patrick Leahy, Eric Davidson, and Gary Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, November 2000.

[160] Yong Peng and Carlo M Croce. The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*, 1:15004, 2016.

[161] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the national academy of sciences*, 98(17):9748–9753, 2001.

[162] Agnieszka Podolska, Christian Anthon, Mads Bak, Niels Tommerup, Kerstin Skovgaard, Peter MH Heegaard, Jan Gorodkin, Susanna Cirera, and Merete Fredholm. Profiling micrornas in lung tissue from pigs infected with *Actinobacillus pleuropneumoniae*. *BMC Genomics*, 13(1):459, 2012.

[163] M. N. Poy, M. Spranger, and M. Stoffel. microRNAs and the regulation of glucose and lipid metabolism. *Diabetes, Obesity & Metabolism*, 9 Suppl 2:67–73, November 2007.

[164] Sanyukta Rana, Shijing Yue, Daniela Stadel, and Margot Zöller. Toward tailored exosomes: the exosomal tetraspanin web contributes to target cell selection. *The international Journal of Biochemistry & Cell Biology*, 44(9):1574–1584, 2012.

[165] BBA Raymond, L Turnbull, C Jenkins, R Madhkoor, I Schleicher, CC Uphoff, CB Whitchurch, M Rohde, and SP Djordjevic. *Mycoplasma hyopneumoniae* resides intracellularly within porcine epithelial cells. *Scientific Reports*, 8(1):17697, 2018.

[166] Benjamin Raymond, Ranya Madhkoor, Ina Schleicher, Cord C Uphoff, Lynne Turnbull, Cynthia B Whitchurch, Manfred Rohde, Matthew P Padula, and Steven P Djordjevic. Extracellular actin is a receptor for *Mycoplasma hyopneumoniae*. *Frontiers in Cellular and Infection Microbiology*, 8:54, 2018.

[167] Narsa M Reddy, Vegiraju Suryanarayana, Dhananjaya V Kalvakolanu, Masayuki Yamamoto, Thomas W Kensler, Paul M Hassoun, Steven R Kleeberger, and Sekhar P Reddy. Innate immunity against bacterial infection following hyperoxia exposure is impaired in nrf2-deficient mice. *The Journal of Immunology*, 183(7):4601–4608, 2009.

[168] Brenda J. Reinhart, Earl G. Weinstein, Matthew W. Rhoades, Bonnie Bartel, and David P. Bartel. MicroRNAs in plants. *Genes & Development*, 16(13):1616–1626, 2002.

[169] Edyta Reszka, Zbigniew Jablonowski, Edyta Wieczorek, Jolanta Gromadzinska, Ewa Jablonska, Marek Sosnowski, and Wojciech Wasowicz. Expression of nrf2 and nrf2-modulated genes in peripheral blood leukocytes of bladder cancer males. *Neoplasma*, 60(2):123–128, 2013.

[170] Peter Rice, Ian Longden, and Alan Bleasby. Emboss: the european molecular biology open software suite, 2000.

[171] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[172] Christine Roden, Jonathan Gaillard, Shaveta Kanoria, William Rennie, Syndi Barish, Jijun Cheng, Wen Pan, Jun Liu, Chris Cotsapas, Ye Ding, and Jun Lu. Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation. *Genome Research*, 27(3):374–384, 2017.

[173] Sophie Rome. Are extracellular micrornas involved in type 2 diabetes and related pathologies? *Clinical Biochemistry*, 46(10-11):937–945, 2013.

[174] J. Graham Ruby, Calvin Jan, Christopher Player, Michael J. Axtell, William Lee, Chad Nusbaum, Hui Ge, and David P. Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, 127(6):1193–1207, 2006.

[175] Santosh B. Satbhai, Daniela Ristova, and Wolfgang Busch. Underground tuning: quantitative regulation of root growth. *Journal of Experimental Botany*, 66(4):1099–1112, 2015.

[176] Jeffrey S Schorey, Yong Cheng, Prachi P Singh, and Victoria L Smith. Exosomes and other extracellular vesicles in host–pathogen interactions. *EMBO Reports*, 16(1):24–43, 2015.

[177] Markus Schueler, Daniela A Braun, Gayathri Chandrasekar, Heon Yung Gee, Timothy D Klasson, Jan Halbritter, Andrea Bieder, Jonathan D Porath, Rannar Airik, Weibin Zhou, et al. Dcdc2 mutations cause a renal-hepatic ciliopathy by disrupting wnt signaling. *The American Journal of Human Genetics*, 96(1):81–92, 2015.

[178] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[179] Cynthia Mira Sharma and Jörg Vogel. Experimental approaches for the discovery and characterization of regulatory small RNA. *Current Opinion in Microbiology*, 12(5):536–546, October 2009.

[180] Lei Shi, France Koll, Olivier Arnaiz, and Jean Cohen. The ciliary protein ift 57 in the macronucleus of paramecium. *Journal of Eukaryotic Microbiology*, 65(1):12–27, 2018.

[181] Liang Shi, Zhan-Guo Chen, Li-li Wu, Jian-Jian Zheng, Jian-Rong Yang, Xiao-Fei Chen, Zeng-Qiang Chen, Cun-Li Liu, Sheng-Ying Chi, Jia-Ying Zheng, et al. mir-340 reverses cisplatin resistance of hepatocellular carcinoma cell lines by targeting nrf2-dependent antioxidant pathway. *Asian Pacific Journal of Cancer Prevention*, 15(23):10439–10444, 2015.

[182] Haim Shirin, Emilia Mia Sordillo, Sung H Oh, Hirofumi Yamamoto, Thomas Delohery, I Bernard Weinstein, and Steven F Moss. *Helicobacter pylori* inhibits the g1 to s transition in ags gastric epithelial cells. *Cancer Research*, 59(10):2277–2281, 1999.

[183] Varsha Shriram, Vinay Kumar, Rachayya M. Devarumath, Tushar S. Khare, and Shabir H. Wani. MicroRNAs As Potential Targets for Abiotic Stress Tolerance in Plants. *Frontiers in Plant Science*, 7, 2016. Publisher: Frontiers.

[184] Katherine J Siddle, Ludovic Tailleux, Matthieu Deschamps, Yong-Hwee Eddie Loh, Cécile Deluen, Brigitte Gicquel, Christophe Antoniewski, Luis B Barreiro, Laurent Farinelli, and Lluis Quintana-Murci. Bacterial infection drives the expression dynamics of micrornas and their isomirs. *PLoS Genetics*, 11(3):e1005064, 2015.

[185] Judith Maxwell Silverman and Neil E Reiner. Exosomes and other microvesicles in infection biology: organelles with unanticipated phenotypes. *Cellular Microbiology*, 13(1):1–9, 2011.

[186] Franciele M Siqueira, Guilherme L de Morais, Susan Higashi, Laura S Beier, Gabriela M Breyer, Caio P de Sá Godinho, Marie-France Sagot, Irene S Schrank, Arnaldo Zaha, and A T R Vasconcelos. Mycoplasma non-coding rna: identification of small rnas and targets. *BMC Genomics*, 17(8):743, 2016.

[187] Martin Sosic and Mile Sikic. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017.

[188] Cathy Staedel and Fabien Darfeuille. Micro rna s and bacterial infection. *Cellular Microbiology*, 15(9):1496–1507, 2013.

[189] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.

[190] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800, 2011.

[191] Aarti Tarkar, Niki T Loges, Christopher E Slagle, Richard Francis, Gerard W Dougherty, Joel V Tamayo, Brett Shook, Marie Cantino, Daniel Schwartz, Charlotte Jahnke, et al. Dyx1c1 is required for axonemal dynein assembly and ciliary motility. *Nature Genetics*, 45(9):995, 2013.

[192] Rajesh K Thimmulappa, Hannah Lee, Tirumalai Rangasamy, Sekhar P Reddy, Masayuki Yamamoto, Thomas W Kensler, and Shyam Biswal. Nrf2 is a critical regulator of the innate immune response and survival during experimental sepsis. *The Journal of Clinical Investigation*, 116(4):984–995, 2016.

[193] Hadi Valadi, Karin Ekström, Apostolos Bossios, Margareta Sjöstrand, James J Lee, and Jan O Lötvall. Exosome-mediated transfer of mrnas and micrornas is a novel mechanism of genetic exchange between cells. *Nature Cell Biology*, 9(6):654, 2007.

[194] Edy M Vilei and Joachim Frey. Genetic and Biochemical Characterization of Glycerol Uptake in *Mycoplasma mycoides* subsp. *mycoides* SC: Its Impact on H2O2Production and Virulence. *Clin. Diagn. Lab. Immunol.*, 8(1):85–92, 2001.

[195] Dimitrios M. Vitsios, Elissavet Kentepozidou, Leonor Quintais, Elia Benito-Gutiérrez, Stijn van Dongen, Matthew P. Davis, and Anton J. Enright. Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Research*, 45(21):e177–e177, 2017.

[196] Pat Wadsworth. Tpx2. *Current Biology*, 25(24):R1156–R1158, 2015.

[197] Ming Wang, Arne Weiberg, Feng-Mao Lin, Bart P. H. J. Thomma, Hsien-Da Huang, and Hailing Jin. Bidirectional cross-kingdom RNAi and fungal uptake of external RNAs confer plant protection. *Nature Plants*, 2(10):16151, 2016.

[198] Won-Jing Wang, Hwee Goon Tay, Rajesh Soni, Geoffrey S Perumal, Mary G Goll, Frank P Macaluso, John M Asara, Jeffrey D Amack, and Meng-Fu Bryan Tsou. CEP162 is an axoneme-recognition protein promoting ciliary transition zone assembly at the cilia base. *Nature Cell Biology*, 15(6):591, 2013.

[199] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*, 31(20):3350–3352, 2015.

[200] Erno Wienholds, Wigard P. Kloosterman, Eric Miska, Ezequiel Alvarez-Saavedra, Eugene Berezikov, Ewart de Bruijn, H. Robert Horvitz, Sakari Kauppinen, and Ronald H. A. Plasterk. MicroRNA expression in zebrafish embryonic development. *Science (New York, N.Y.)*, 309(5732):310–311, July 2005.

[201] Patrick Wolter, Kathrin Schmitt, Marc Fackler, Heidi Kremling, Leona Probst, Stefanie Hauser, Oliver J Gruss, and Stefan Gaubatz. GAS2L3, a target gene of the DREAM complex, is required for proper cytokinesis and genomic stability. *J Cell Sci*, 125(10):2393–2406, 2012.

[202] Sarah Woolner, Lori L O'Brien, Christiane Wiese, and William M Bement. Myosin-10 and actin filaments are essential for mitotic spindle function. *J Cell Biol*, 182(1):77–88, 2008.

[203] Huangming Xie, Lei Sun, and Harvey F. Lodish. Targeting microRNAs in obesity. *Expert Opinion on Therapeutic Targets*, 13(10):1227–1238, October 2009.

[204] Liming Xie, Zhiwei Zhang, Zhiqin Tan, Rongfang He, Xi Zeng, Yuanjie Xie, Suyun Li, Guohua Tang, Hailin Tang, and Xiusheng He. MicroRNA-124 inhibits proliferation and induces apoptosis by directly repressing EZH2 in gastric cancer. *Molecular and Cellular Biochemistry*, 392(1-2):153–159, July 2014.

[205] Jiahong Xu, Yu Tang, Yihua Bei, Shengguang Ding, Lin Che, Jianhua Yao, Hongbao Wang, Dongchao Lv, and Junjie Xiao. miR-19b attenuates H2O2-induced apoptosis in rat H9C2 cardiomyocytes via targeting PTEN. *Oncotarget*, 7(10):10870, 2016.

[206] Junfei Xu, Feiran Wang, Xi Wang, Zhixian He, and Xinguo Zhu. miRNA-543 promotes cell migration and invasion by targeting SPOP in gastric cancer. *OncoTargets and therapy*, 11:5075–5082, August 2018.

[207] Xiao-Ling Xu, Wei Ma, Yu-Bo Zhu, Chao Wang, Bing-Yuan Wang, Na An, Lei An, Yan Liu, Zhong-Hong Wu, and Jian-Hui Tian. The microtubule-associated protein ASPM regulates spindle assembly and meiotic progression in mouse oocytes. *PLoS One*, 7(11):e49303, 2012.

[208] F Xue, H Li, J Zhang, J Lu, Y Xia, and Q Xia. miR-31 regulates interleukin 2 and kinase suppressor of ras 2 during T cell activation. *Genes and Immunity*, 14(2):127, 2013.

[209] Xiaochang Xue, Anthony T Cao, Xiaocang Cao, Suxia Yao, Eric D Carlsen, Lynn Soong, Chang-Gong Liu, Xiuping Liu, Zhanju Liu, L Wayne Duck, et al. Downregulation of micro RNA-107 in intestinal CD 11c+ myeloid cells in response to microbiota and proinflammatory cytokines increases IL-23p19 expression. *European Journal of Immunology*, 44(3):673–682, 2014.

[210] Heping Yang, Tony WH Li, Yu Zhou, Hui Peng, Ting Liu, Ebrahim Zandi, María Luz Martínez-Chantar, José M Mato, and Shelly C Lu. Activation of a novel c-Myc-miR27-prohibitin 1 circuitry in cholestatic liver injury inhibits glutathione synthesis in mice. *Antioxidants & Redox Signaling*, 22(3):259–274, 2015.

[211] Muhua Yang, Yuan Yao, Gabriel Eades, Yongshu Zhang, and Qun Zhou. MiR-28 regulates Nrf2 expression through a Keap1-independent mechanism. *Breast Cancer Research and Treatment*, 129(3):983–991, 2011.

[212] Xiaozeng Yang and Lei Li. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*, 27(18):2614–2615, 2011.

[213] Takayuki Yasunaga, Sylvia Hoff, Christoph Schell, Martin Helmstädter, Oliver Kretz, Sebastian Kuechlin, Toma A Yakulov, Christina Engel, Barbara Müller, Robert Bensch, et al. The polarity protein inturned links nphp4 to daam1 to control the subapical actin network in multiciliated cells. *J Cell Biol*, 211(5):963–973, 2015.

[214] Theresa F Young, Eileen L Thacker, Barbara Z Erickson, and Richard F Ross. A tissue culture system to study respiratory ciliary epithelial adherence of selected swine mycoplasmas. *Veterinary Microbiology*, 71(3-4):269–279, 2000.

[215] Baohong Zhang, Xiaoping Pan, George P. Cobb, and Todd A. Anderson. microRNAs as oncogenes and tumor suppressors. *Developmental Biology*, 302(1):1–12, February 2007.

[216] Yong-Mei Zhang, Jennifer M. Noto, Charles E. Hammond, Jeremy L. Barth, W. Scott Argraves, Steffen Backert, Richard M. Peek, and Adam J. Smolka. *Helicobacter pylori*-induced posttranscriptional regulation of H-K-ATPase Œ-subunit gene expression by miRNA. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 306(7):G606–G613, April 2014.

[217] Lin Zheng, Eric Leung, Nelson Lee, Grace Lui, Ka-Fai To, Raphael CY Chan, and Margaret Ip. Differential microRNA expression in human macrophages with *Mycobacterium tuberculosis* infection of Beijing/W and non-Beijing/W strain types. *PLoS One*, 10(6):e0126018, 2015.

[218] Jian Zhou, Fang Yang, N Adrian Leu, and P Jeremy Wang. MNS1 is essential for spermiogenesis and motile ciliary functions in mice. *PLoS Genetics*, 8(3):e1002516, 2012.

[219] Xikun Zhou, Xuefeng Li, and Min Wu. miRNAs reshape immunity and inflammatory responses in bacterial infection. *Signal Transduction and Targeted Therapy*, 3:14, 2018.

[220] GC Zielinski and RF Ross. Effect of growth in cell cultures and strain on virulence of *Mycoplasma hyopneumoniae* for swine. *American Journal of Veterinary Research*, 51(3):344–348, 1990.

[221] Mark Ziemann, Antony Kaspi, and Assam El-Osta. Evaluation of microRNA alignment techniques. *RNA*, 22(8):1120–1138, 2016.

[222] Fernanda M. A. L. Zimmer, Gabriela P. Paludo, Hercules Moura, John R. Barr, and Henrique B. Ferreira. Differential secretome profiling of a swine tracheal cell line infected with mycoplasmas of the swine respiratory tract. *Journal of Proteomics*, 192:147–159, 2019.

Titre : Développement de nouveaux algorithmes pour avancer dans la découverte des microARNs

Résumé : Les microARNs (miARNs) sont de petits ARNs non codants qui jouent un rôle clé dans la régulation de l'expression génique. Les miARNs sont impliqués dans un large éventail de processus biologiques, notamment le cycle cellulaire, la différenciation, l'apoptose et les maladies. Au cours de la dernière décennie, avec l'accessibilité croissante des technologies de séquençage à haut débit, les expériences de sARN-seq ont permis d'identifier les miARNs et de prédire le réseau de régulation miARN-ARNm pour mieux comprendre leur rôle et leur fonction biologiques. Différents algorithmes ont été développés au cours des dernières années pour y parvenir, mais il s'est avéré difficile de réaliser une caractérisation complète des miARNs et de leurs cibles qui permettrait de bien les identifier et les annoter. Les travaux de cette thèse ont débuté par une participation à l'analyse expérimentale et bioinformatique de données dual miARN-seq et mARN-seq obtenues en profilant l'interaction hôte-pathogène de *Sus scrofa* avec la bactérie *Mycoplasma hyopneumoniae*. L'expérience que j'ai ainsi acquise avec les outils de pointe pour la découverte de miARNs et la prédiction de leurs cibles a été essentielle pour identifier les faiblesses des algorithmes actuels et donc le réel besoin de développer de nouveaux algorithmes liés à la première étape de l'analyse de miARNs, à savoir leur identification. Cela a représenté alors l'objectif principal de cette thèse. Avec cet objectif à l'esprit, j'ai développé la boîte à outils BrumiR, qui est un package composé de trois parties : 1) un nouvel outil de découverte de miRNAs appelé BrumiR-core, 2) un algorithme spécifique d'alignement sur génome appelé BrumiR2Reference, et 3) un simulateur de lecture sARN-seq appelé miRsim. En particulier, BrumiR-core est un algorithme *de novo* basé sur une approche de graphe de de Bruijn qui est capable d'identifier les miARNs directement et exclusivement à partir de données de sARN-seq. Nous avons effectué un benchmark de BrumiR en utilisant des jeux de données simulées et réelles de sRNA-seq d'espèces animales et végétales. Les résultats montrent que BrumiR atteint le rappel le plus élevé pour la découverte de miARNs, tout en étant beaucoup plus rapide et plus efficace que les outils de pointe évalués. En résumé, nous présentons une nouvelle méthode polyvalente qui met en œuvre de nouvelles idées algorithmiques pour l'étude des miARNs qui complète et étend les approches actuellement existantes. Le code de la boîte à outils BrumiR est disponible gratuitement dans GitHub (https://github.com/camoragaq).

Mots-clés : sARN; miARNs; découverte *de novo* de miARNs; NGS; sRNA-seq; graphe de Bruijn; algorithmes.

Title: Development of new algorithms to advance on the discovery of microRNAs

Abstract: MicroRNAs (miRNAs) are small non-coding RNAs that are key players in the regulation of gene expression. miRNAs are involved in a wide range of biological processes including cell cycle, differentiation, apoptosis, and disease. In the last decade, with the increasing accessibility of high-throughput sequencing technologies, sRNAs-seq experiments have provided the opportunity to identify miRNAs, and to predict the miRNA-mRNA regulatory network to better understand their biological role and function. Different algorithms have been developed during the last years to achieve this, but it has proven difficult to achieve a complete characterization of miRNAs and of their targets that would enable to well identify and annotate them. The work in this thesis started by a participation in the experimental and bioinformatic analysis of dual miRNA-seq and mRNA-seq data obtained by profiling the host-pathogen interaction of *Sus scrofa* with the bacterium *Mycoplasma hyopneumoniae*. The experience I thus acquired with the current state-of-the-art tools for miRNA discovery and miRNA target prediction was essential to identify the weaknesses of the current tools and therefore the real need to develop new algorithms related to the first step of the analysis of miRNAs, namely their identification. This then represented the main objective of this thesis. With that objective in mind, I developed the BrumiR toolkit, which is a package composed of three tools: 1) a new discovery miRNA tool called BrumiR-core, 2) a specific genome mapper called BrumiR2Reference, and 3) a sRNA-seq read simulator called miRsim. In particular, BrumiR-core is a *de novo* algorithm based on a de Bruijn graph approach that is able to identify miRNAs directly and exclusively from sRNA-seq data. We benchmarked BrumiR using simulated and real sRNA-seq data of both animal and plant species. The results demonstrate that BrumiR reaches the highest recall for miRNA discovery, while at the same time being much faster and more efficient than the state-of-the-art tools evaluated. In summary, we present a new and versatile method that implements novel algorithmic ideas for the study of miRNAs that complements and extends the currently existing approaches. The code of the BrumiR toolkit is freely available in GitHub (https://github.com/camoragaq).

Keywords: sRNA; miRNAs; *de novo* miRNA discovery; NGS; sRNA-seq; de Bruijn graph; algorithms.